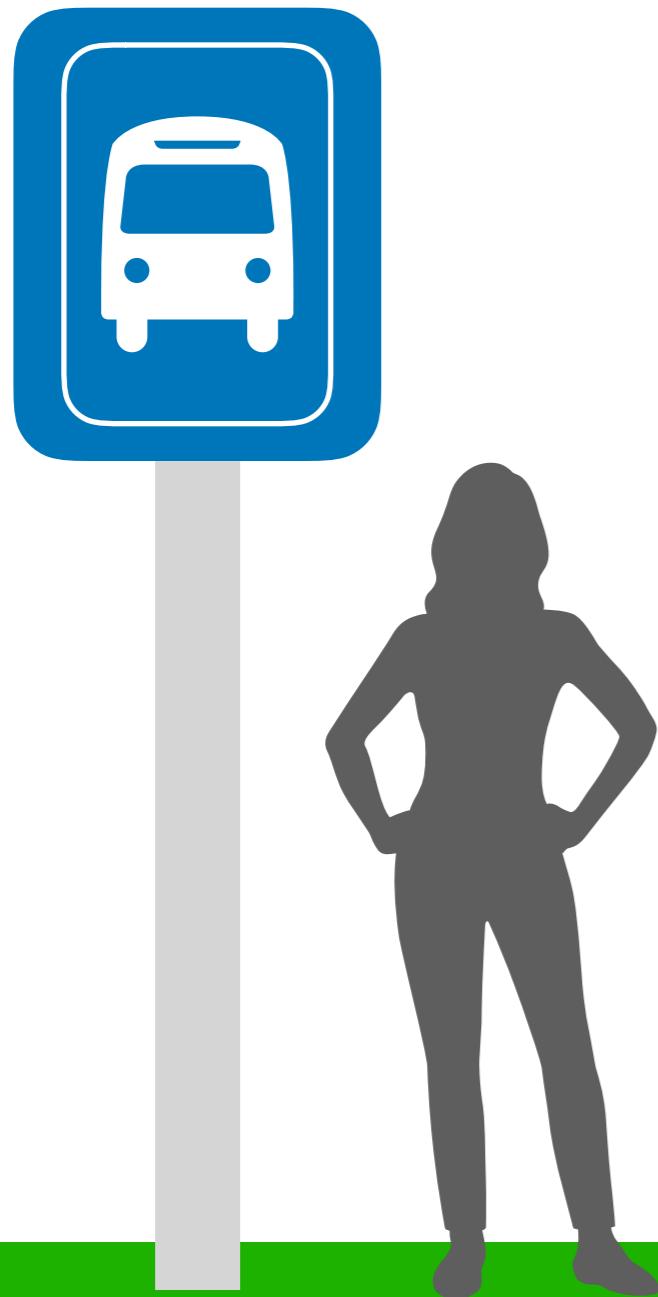


# Lecture 12

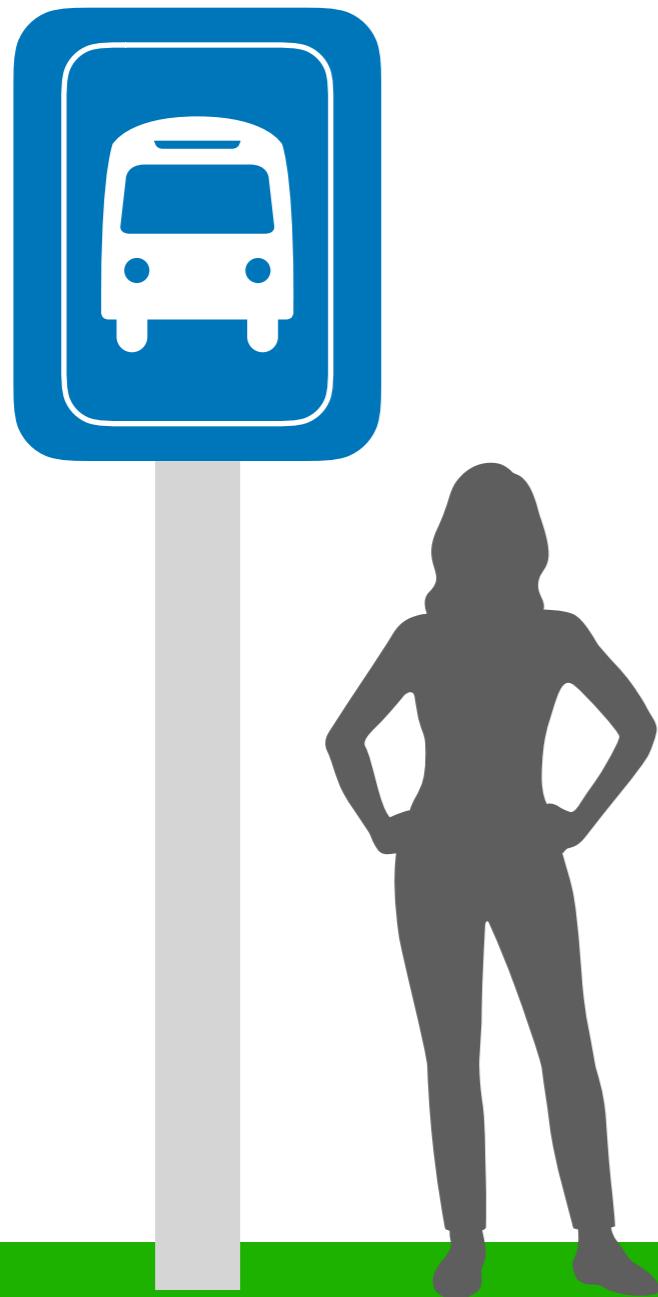
Bayesian inference  
Botany 563 – Spring 2021

- **Previous class check-up:**
  - We reviewed the distance, parsimony and likelihood methods and their strengths and weaknesses
- **Learning Objectives:** At the end of today's session, you will be able to
  - Explain the main characteristics of Bayesian inference for phylogenetics
  - Understand the role that priors, sample size and convergence play in the performance of Bayesian inference
  - Learn how to use PhyloBayes and/or MrBayes
- **Pre-class work**
  - Read HAL 1.4 and Nascimento et al (2017)

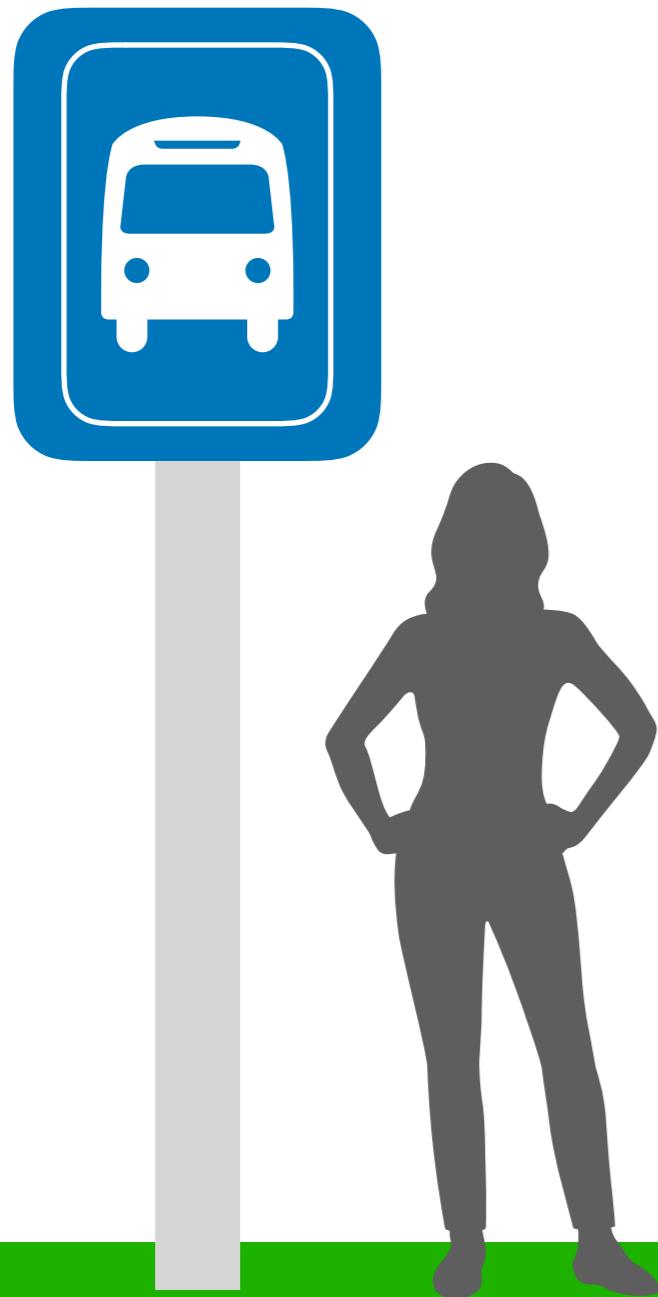
# A familiar example



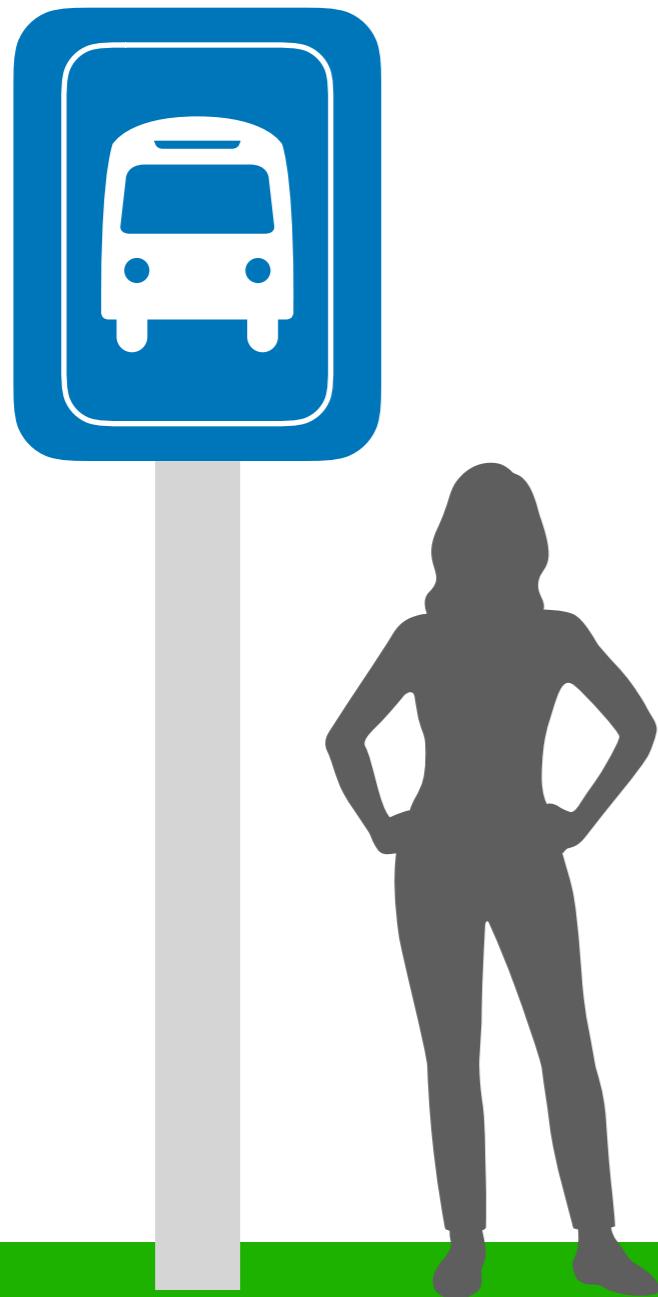
# A familiar example



# A familiar example



# A familiar example



# A familiar example

$X_i = \text{Number of bikes seen on day } i$

$$X_i \sim \text{Poisson}(\lambda)$$

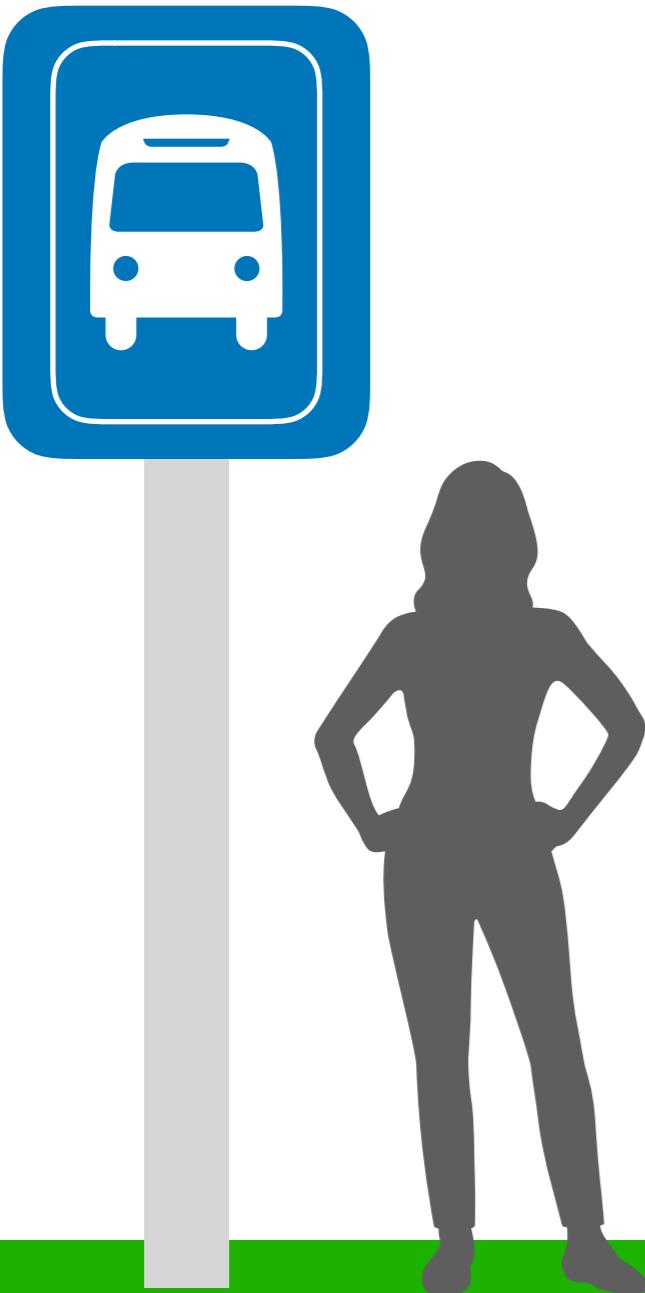
$$P(X_i = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

We collect data on  $n$  days:

$$X_1, X_2, \dots, X_n$$

**We want the MLE for  $\lambda$**

1. We need the likelihood function
2. We need to maximize the likelihood function



# A familiar example

$X_i = \text{Number of bikes seen on day } i$

$$X_i \sim \text{Poisson}(\lambda)$$

$$P(X_i = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

We collect data on  $n$  days:

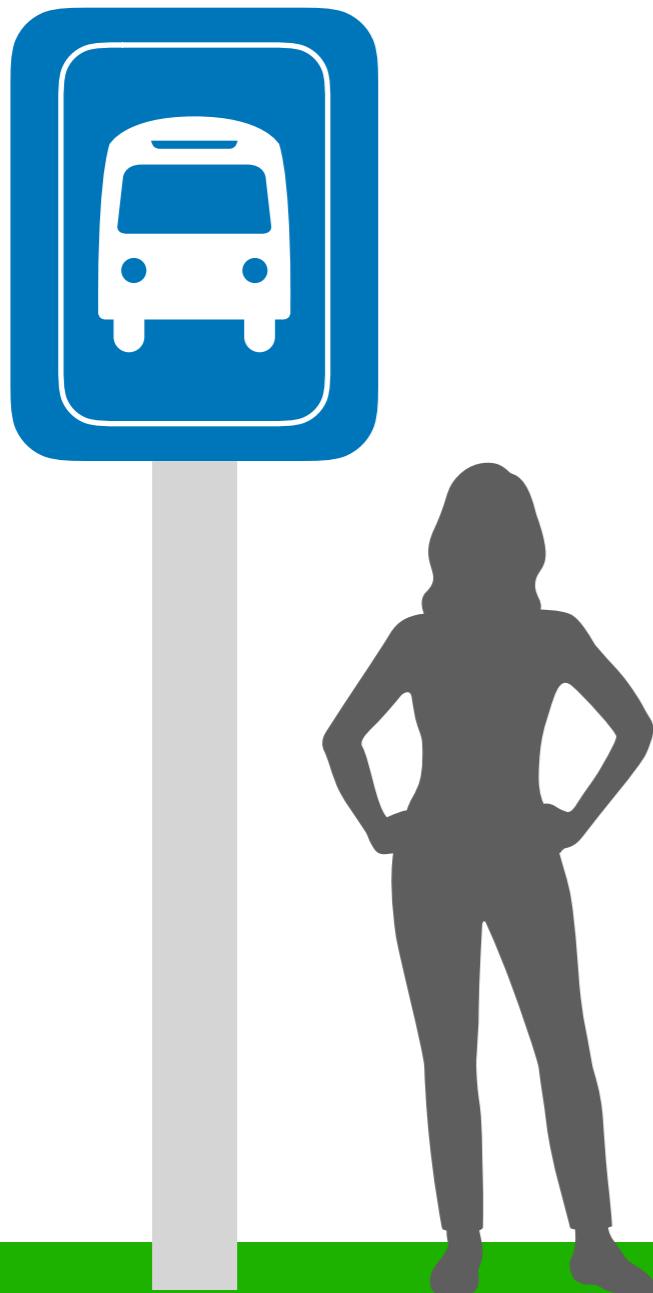
$$X_1, X_2, \dots, X_n$$

Sequences

**We want the MLE for  $\lambda$**

Tree, BL, Q

1. We need the likelihood function
2. We need to maximize the likelihood function



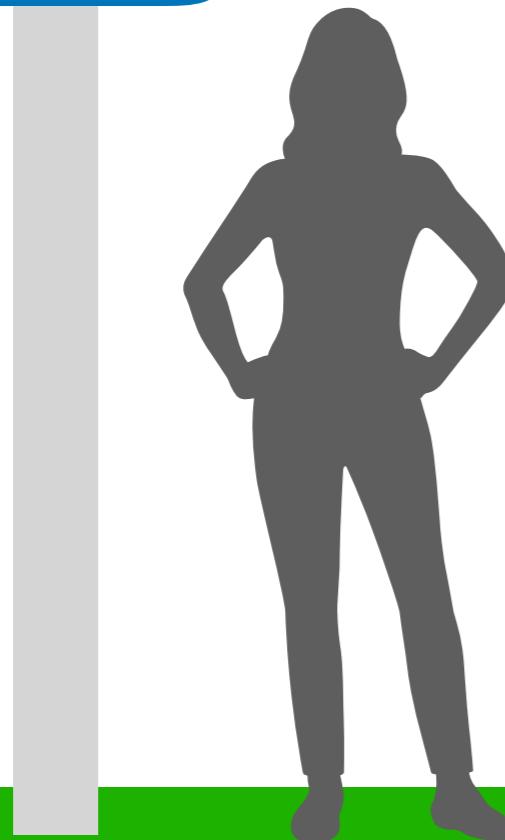
# A familiar example

$X_i = \text{Number of bikes seen on day } i$



$$X_i \sim \text{Poisson}(\lambda) \quad P(X_i = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

We collect data on  $n$  days:  $X_1, X_2, \dots, X_n$



# A familiar example

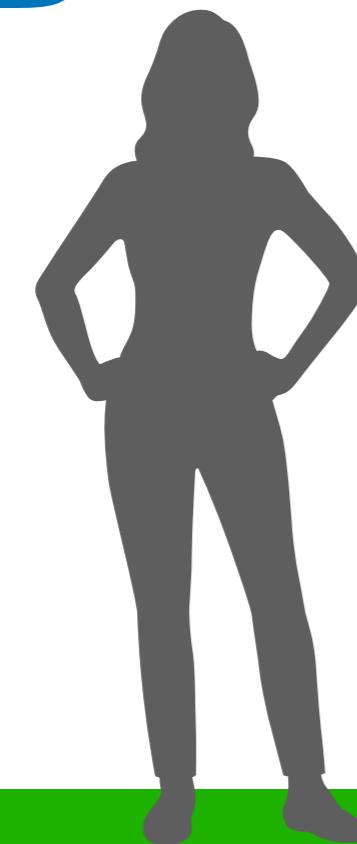
$X_i$  = Number of bikes seen on day  $i$



$$X_i \sim Poisson(\lambda) \quad P(X_i = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

We collect data on  $n$  days:  $X_1, X_2, \dots, X_n$

$$L(\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!}$$



# A familiar example

$X_i = \text{Number of bikes seen on day } i$



$$X_i \sim \text{Poisson}(\lambda) \quad P(X_i = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

We collect data on  $n$  days:  $X_1, X_2, \dots, X_n$

$$L(\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!}$$

$$l(\lambda) = -n\lambda + (\sum x_i) \log(\lambda) - \log(\prod x_i!)$$



# A familiar example

$X_i = \text{Number of bikes seen on day } i$



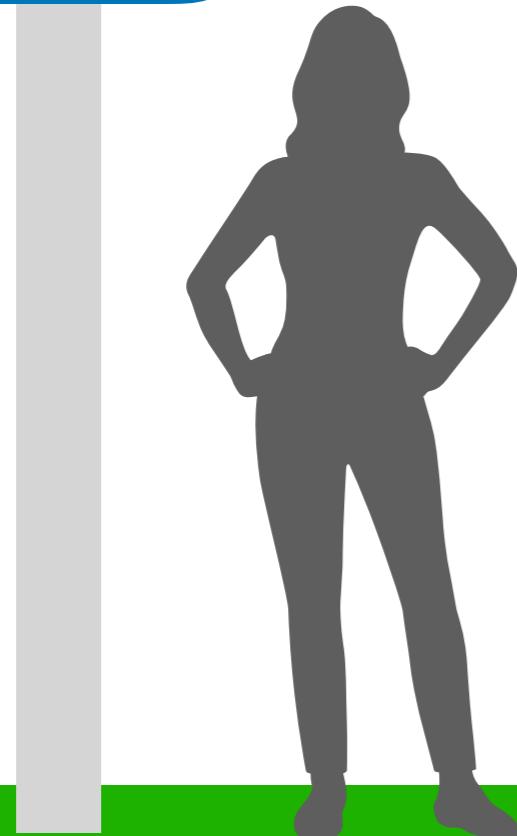
$$X_i \sim \text{Poisson}(\lambda) \quad P(X_i = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

We collect data on  $n$  days:  $X_1, X_2, \dots, X_n$

$$L(\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!}$$

$$l(\lambda) = -n\lambda + (\sum x_i) \log(\lambda) - \log(\prod x_i!)$$

$$l'(\lambda) = -n + \frac{\sum x_i}{\lambda} = 0$$



# A familiar example

$X_i$  = Number of bikes seen on day  $i$



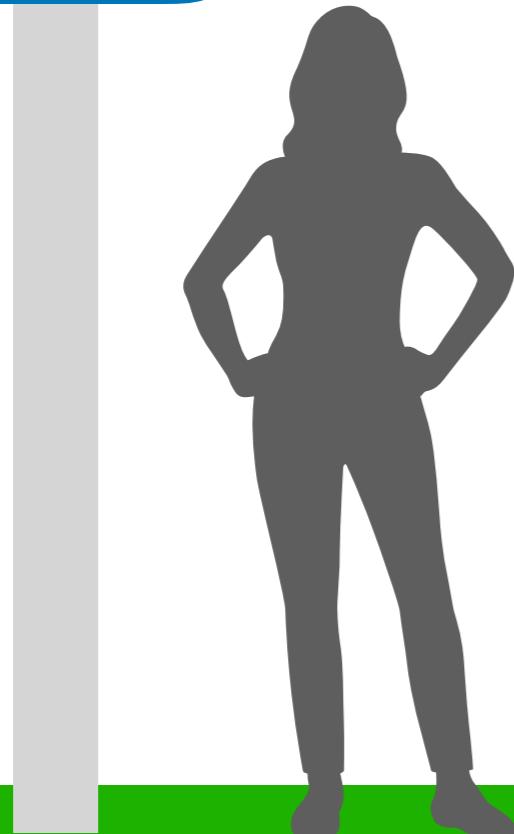
$$X_i \sim \text{Poisson}(\lambda) \quad P(X_i = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

We collect data on  $n$  days:  $X_1, X_2, \dots, X_n$

$$L(\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!}$$

$$l(\lambda) = -n\lambda + (\sum x_i) \log(\lambda) - \log(\prod x_i!)$$

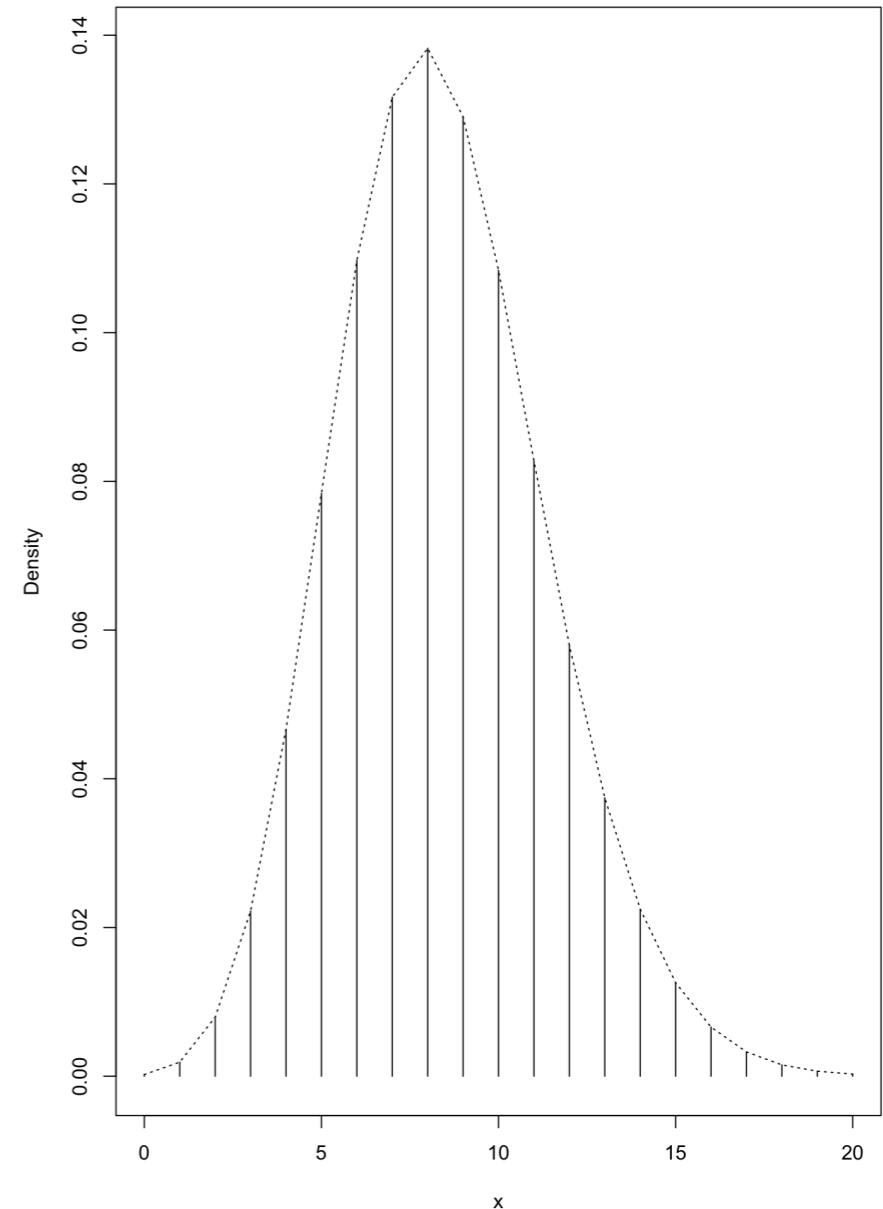
$$l'(\lambda) = -n + \frac{\sum x_i}{\lambda} = 0 \rightarrow \hat{\lambda} = \frac{\sum x_i}{n}$$



# Bayesian: Likelihood 2.0

Information in the data

Likelihood

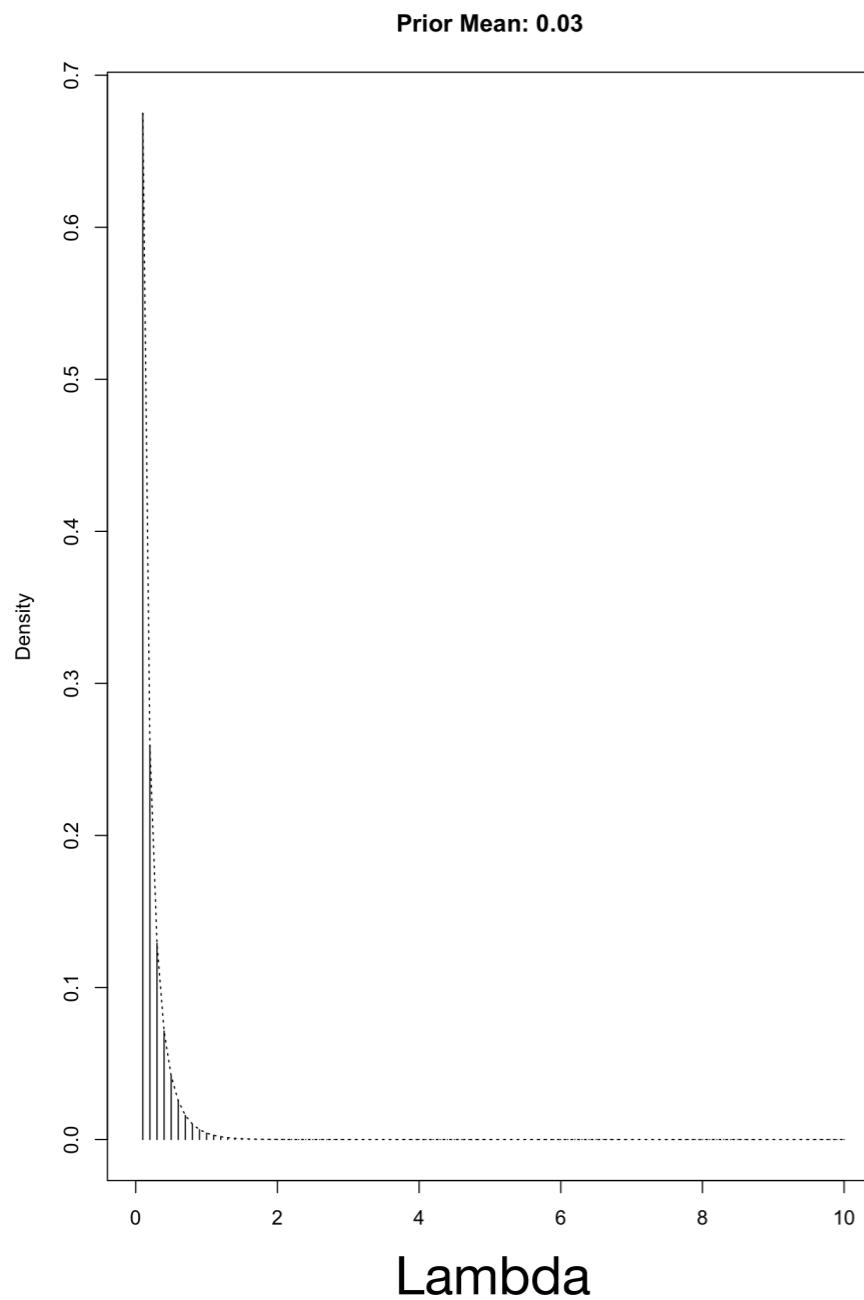


$$\hat{\lambda} = 8.4$$

# Bayesian: Likelihood 2.0

Your knowledge

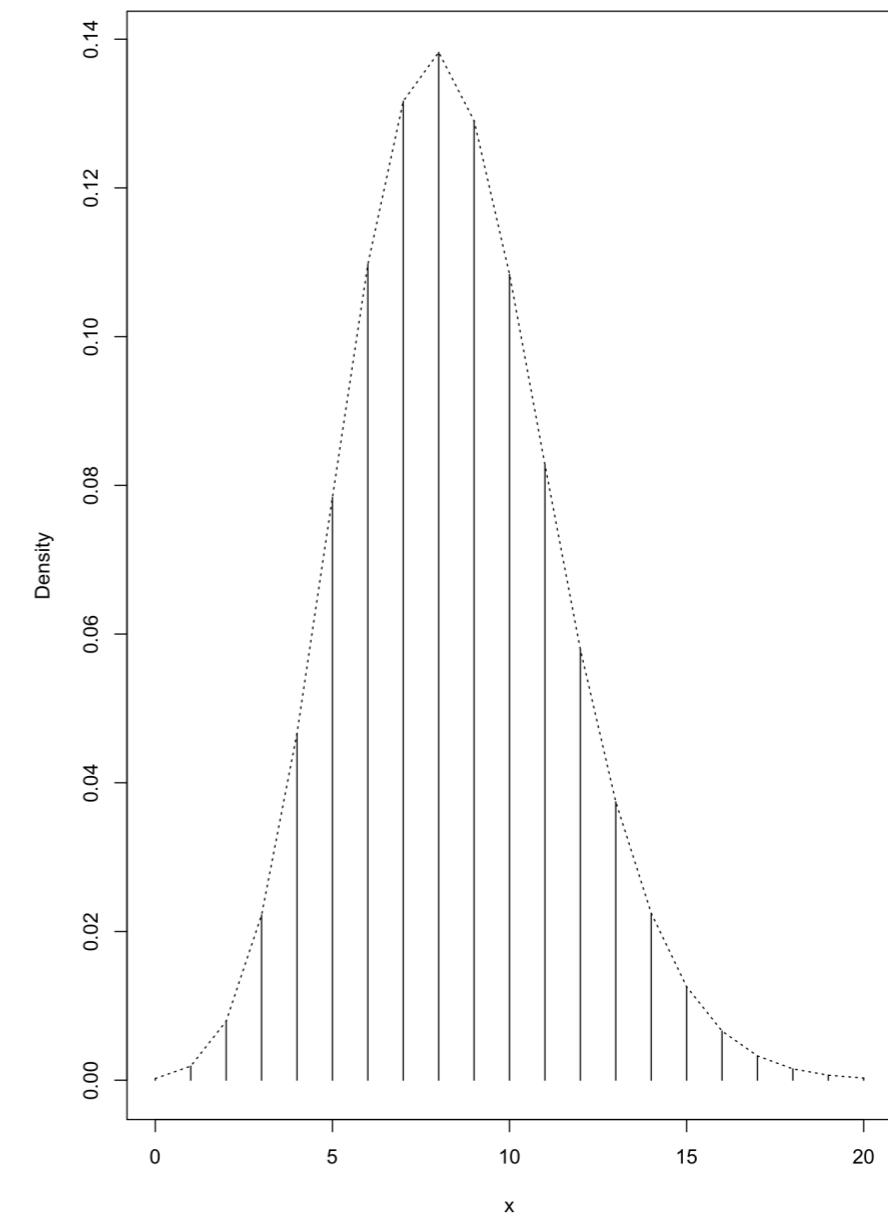
Prior



$$\lambda \sim \text{Gamma}(\alpha = 0.1, \beta = 3.3)$$

Information in the data

Likelihood



$$\hat{\lambda} = 8.4$$

# Bayesian: Likelihood 2.0

Your knowledge

Prior

Information in the data

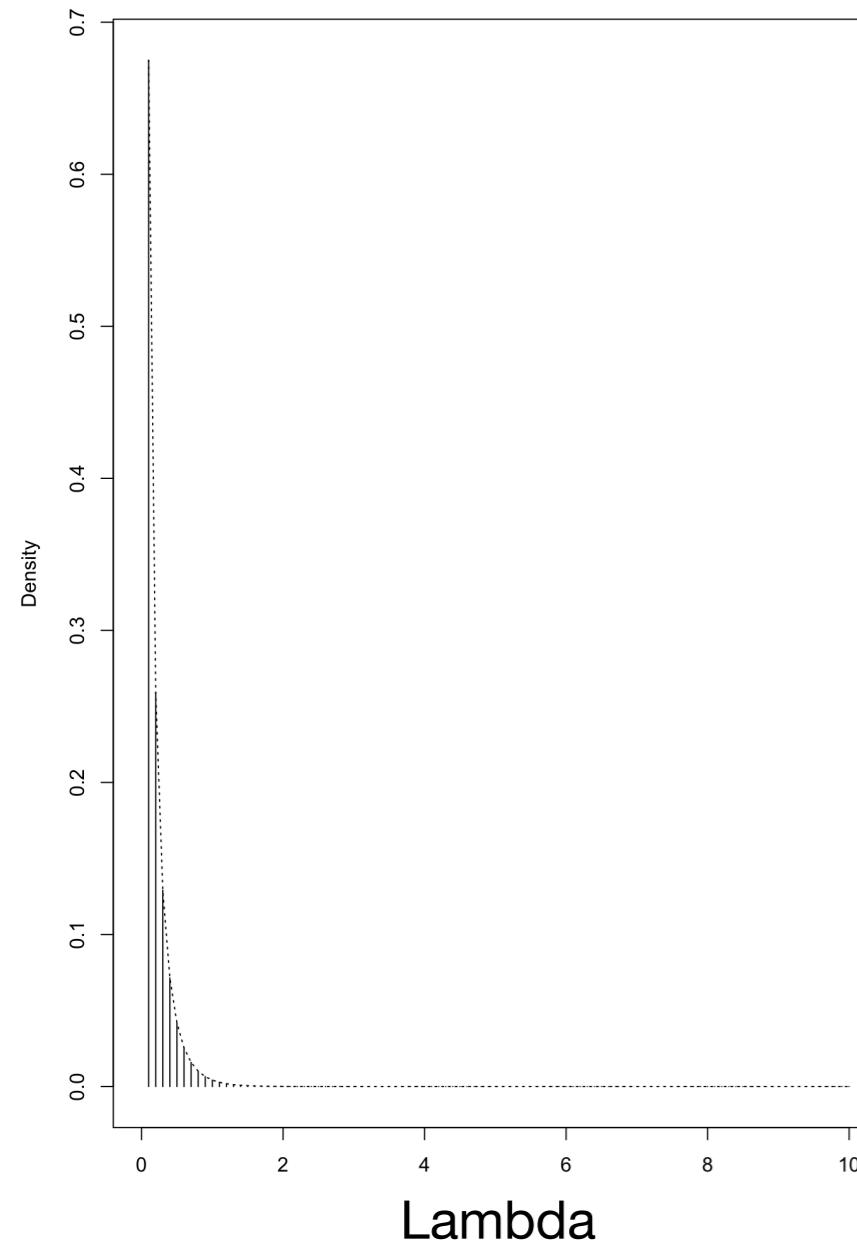
Likelihood

Inference on parameter

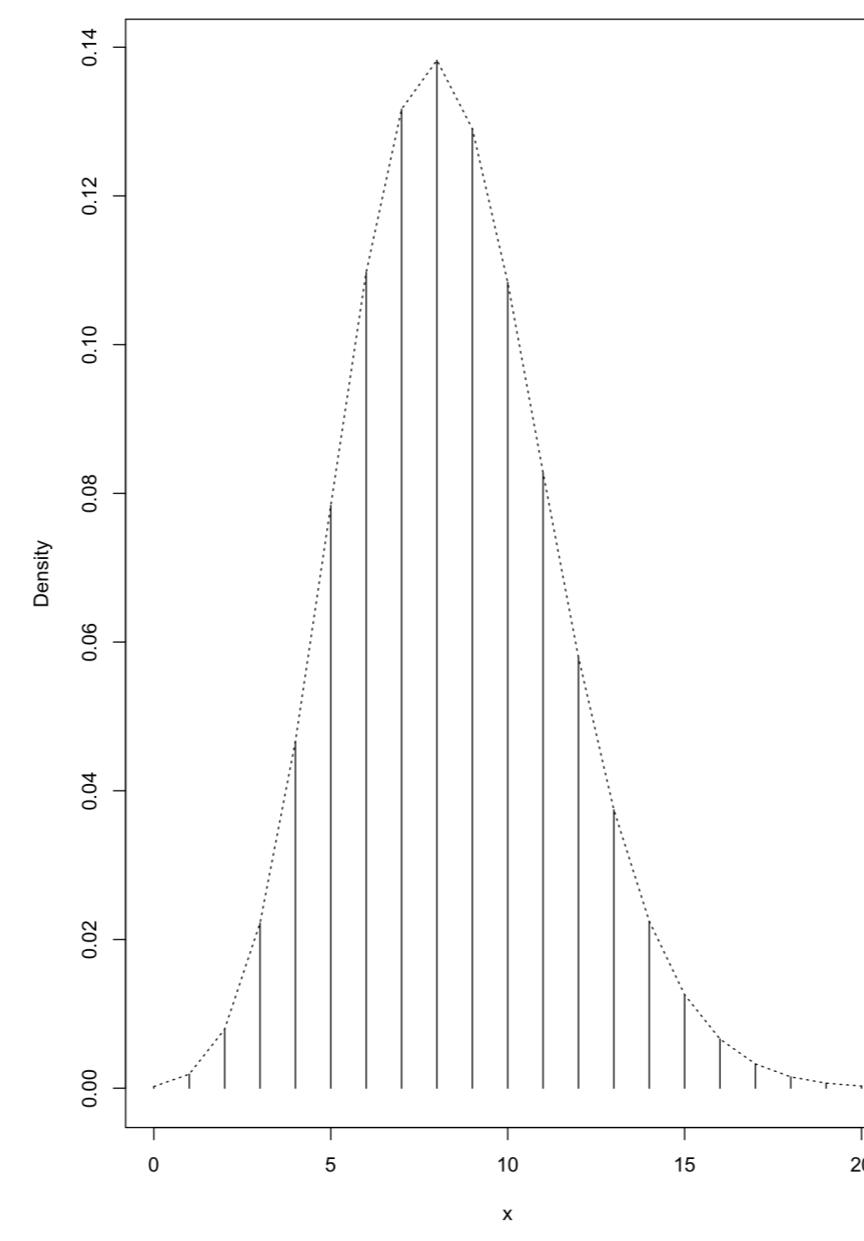
Posterior

Prior Mean: 0.03

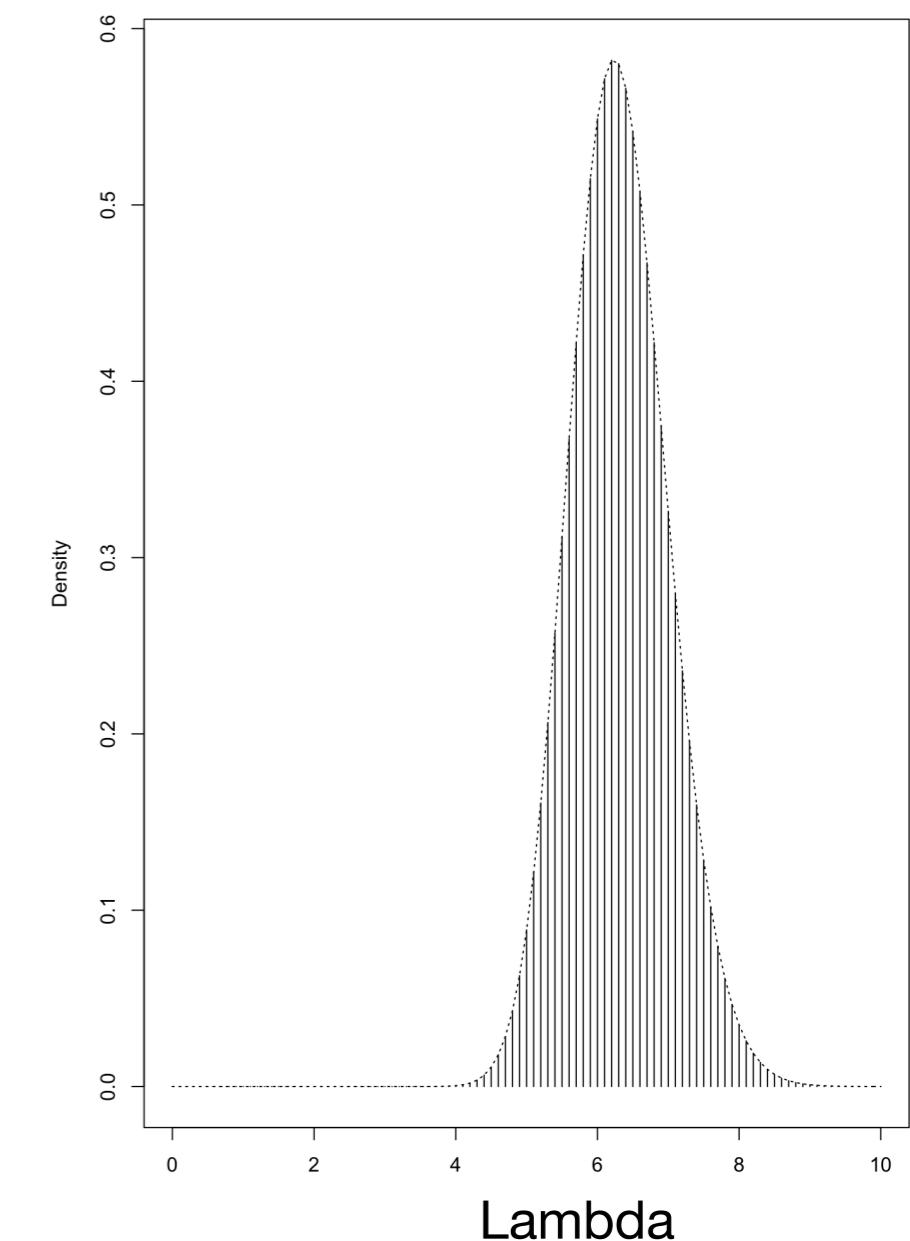
Posterior Mean: 6.31



$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$



$$\hat{\lambda} = 8.4$$



$\lambda | \mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$

# Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

Your knowledge

Prior

Prior Mean: 0.03

Information in the data

Likelihood

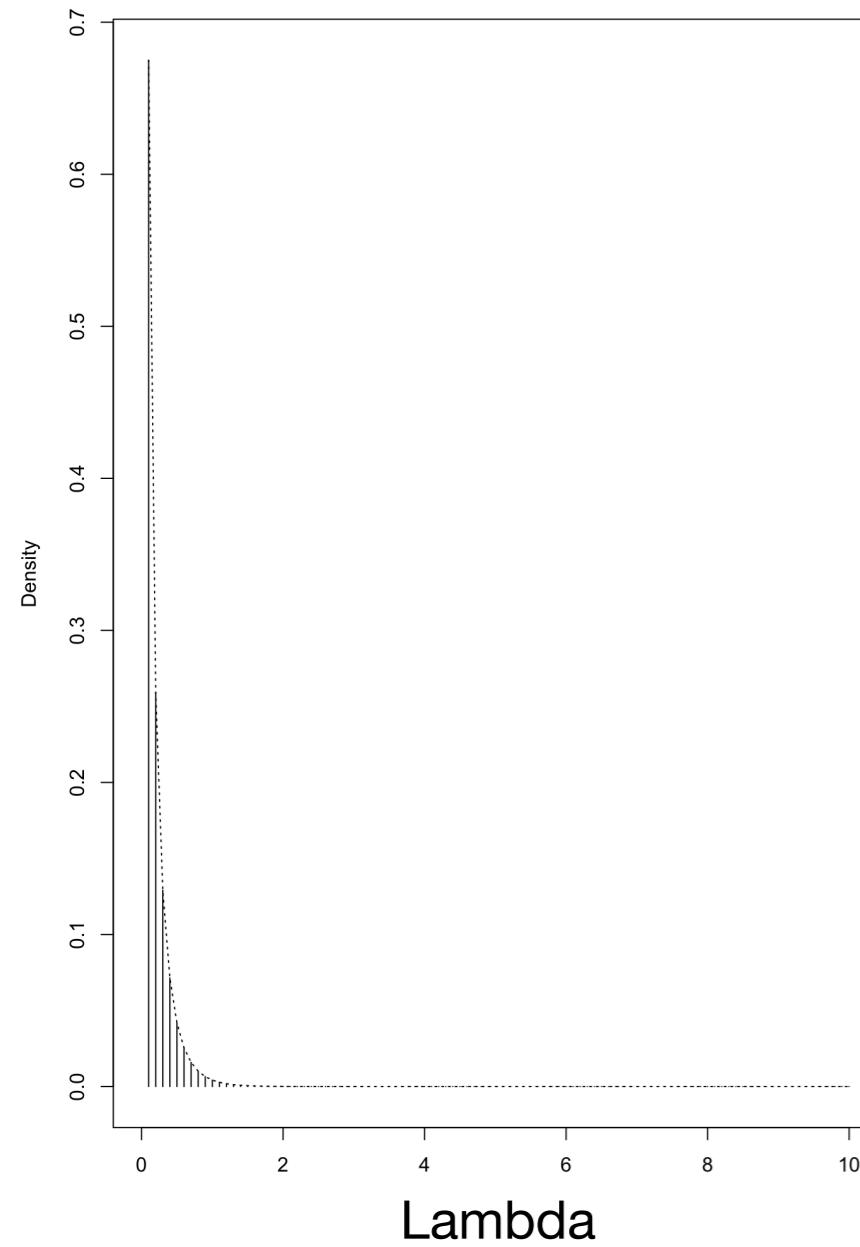
+

Inference on parameter

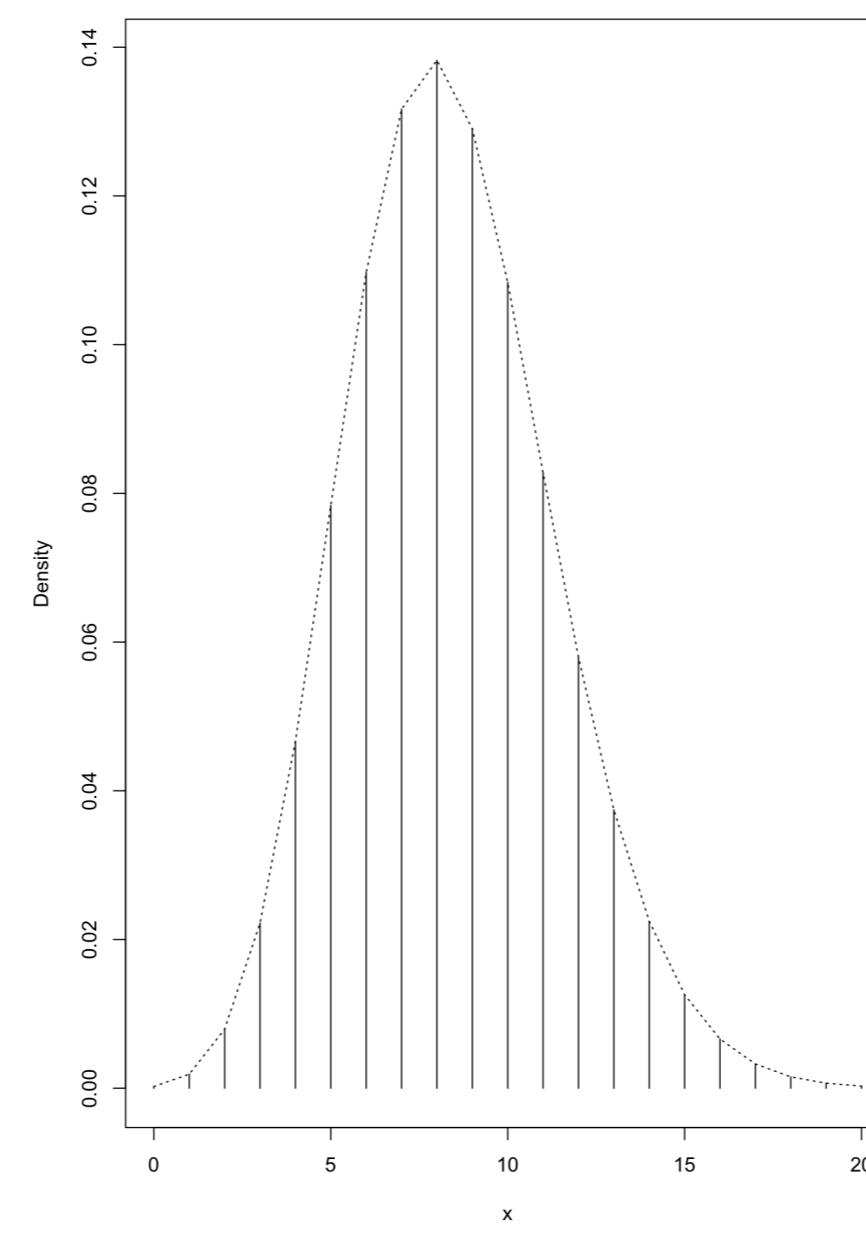
Posterior

=

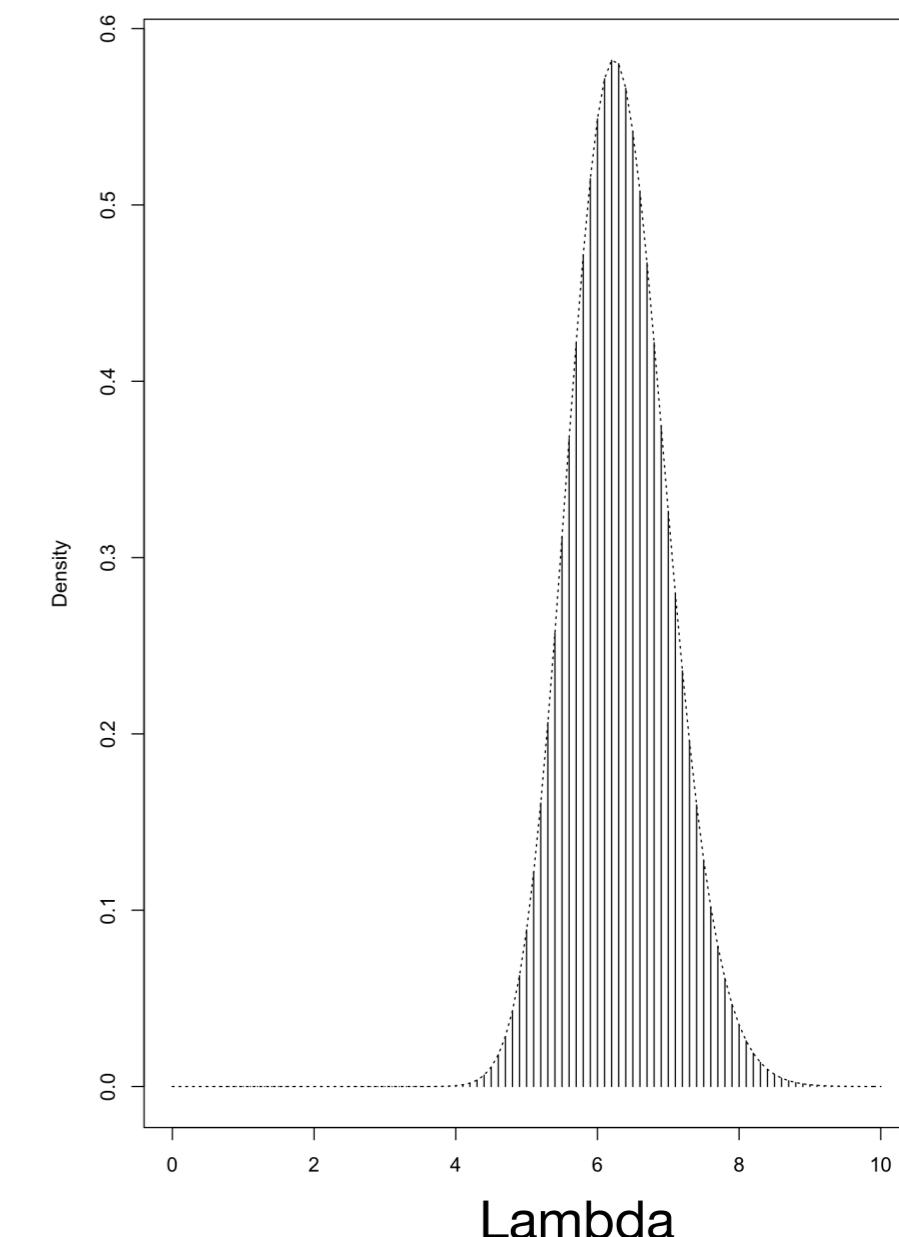
Posterior Mean: 6.31



$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$



$$\hat{\lambda} = 8.4$$



$\lambda | \mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$

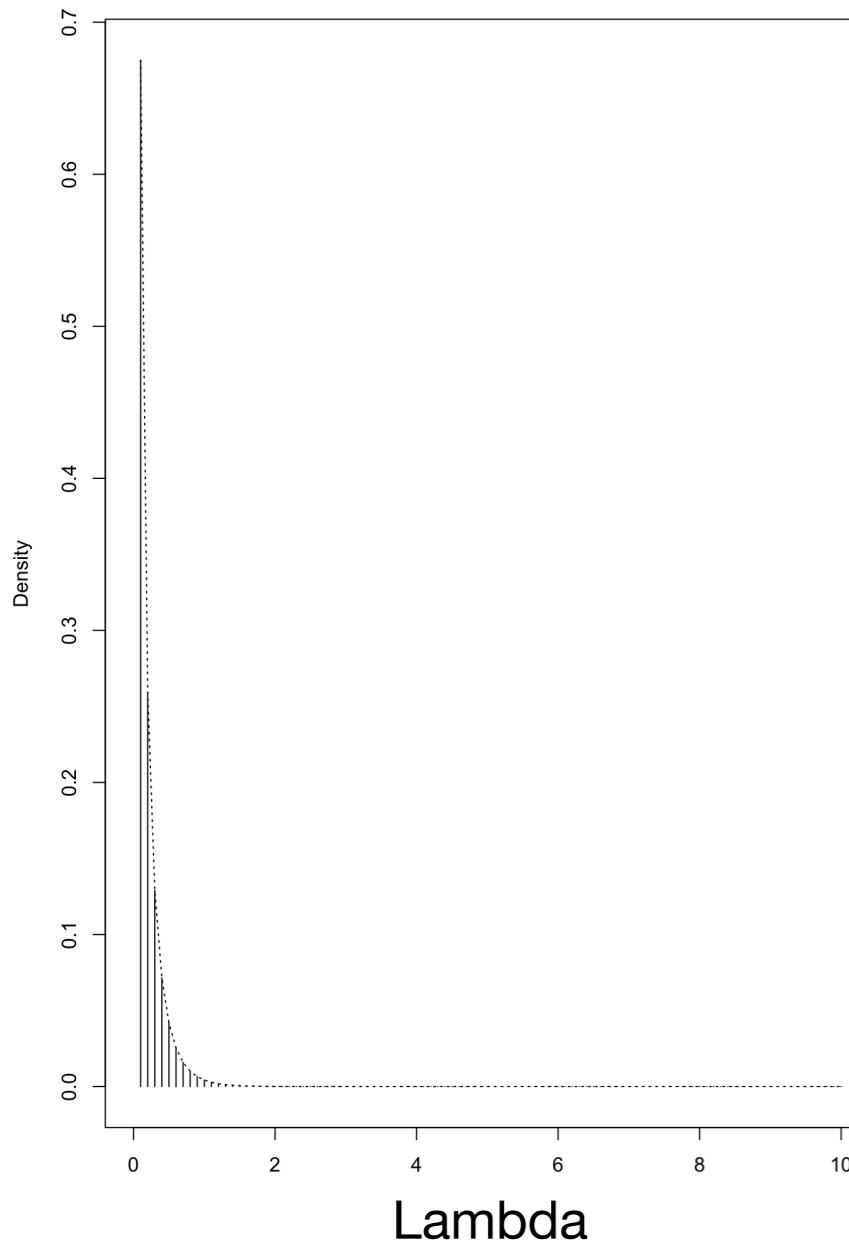
# Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

Your knowledge

Prior

Prior Mean: 0.03

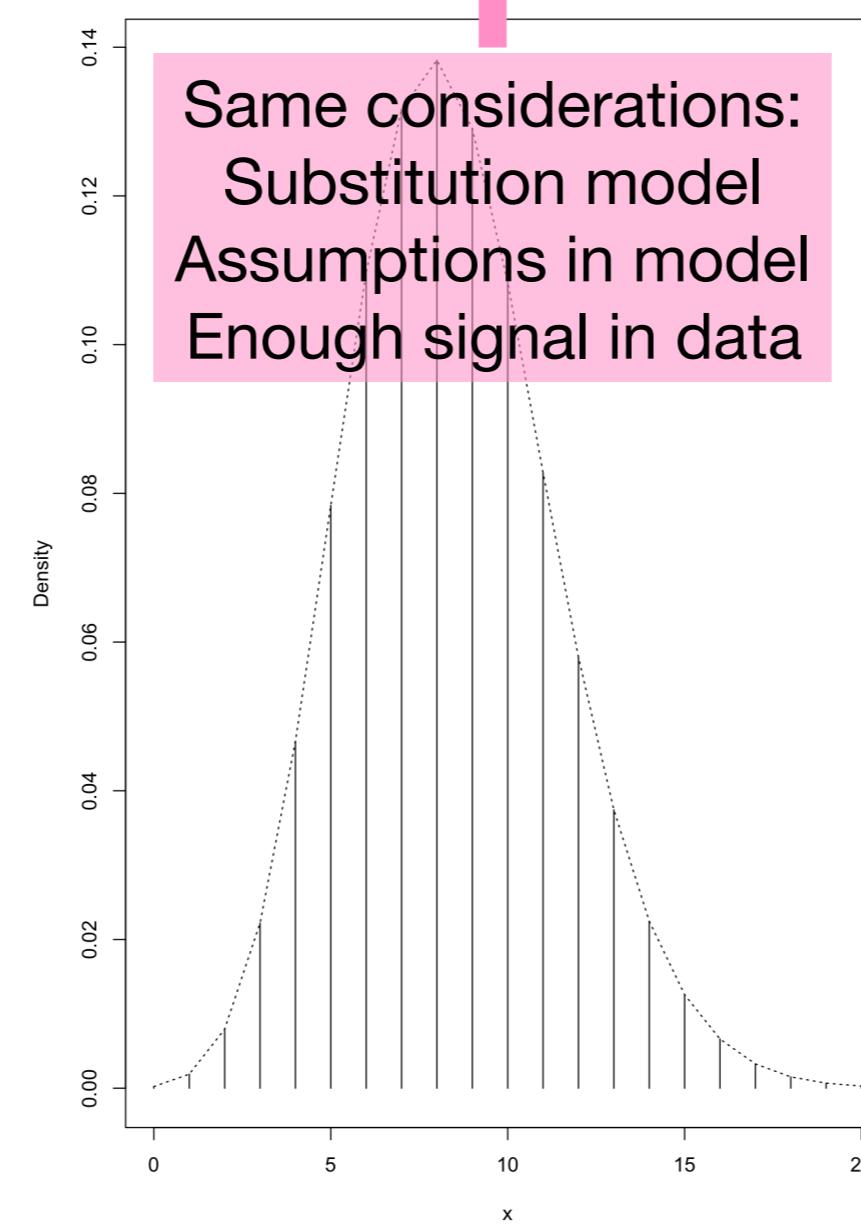


Information in the data

Likelihood



Same considerations:  
Substitution model  
Assumptions in model  
Enough signal in data



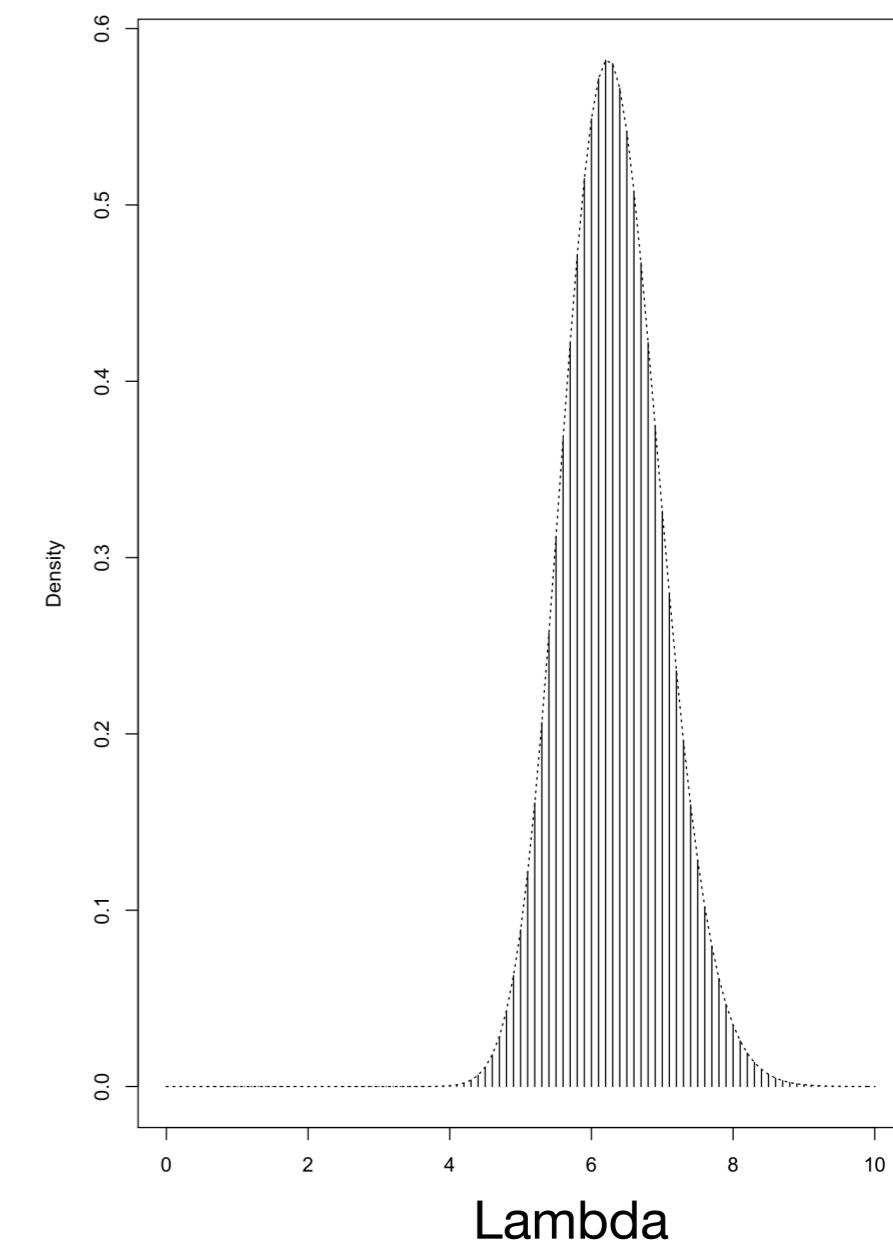
$$\lambda \sim \text{Gamma}(\alpha = 0.1, \beta = 3.3)$$

$$\hat{\lambda} = 8.4$$

Inference on parameter

Posterior

Posterior Mean: 6.31



$$\lambda | \mathbf{X} \sim \text{Gamma}(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$

# Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

Your knowledge

Prior

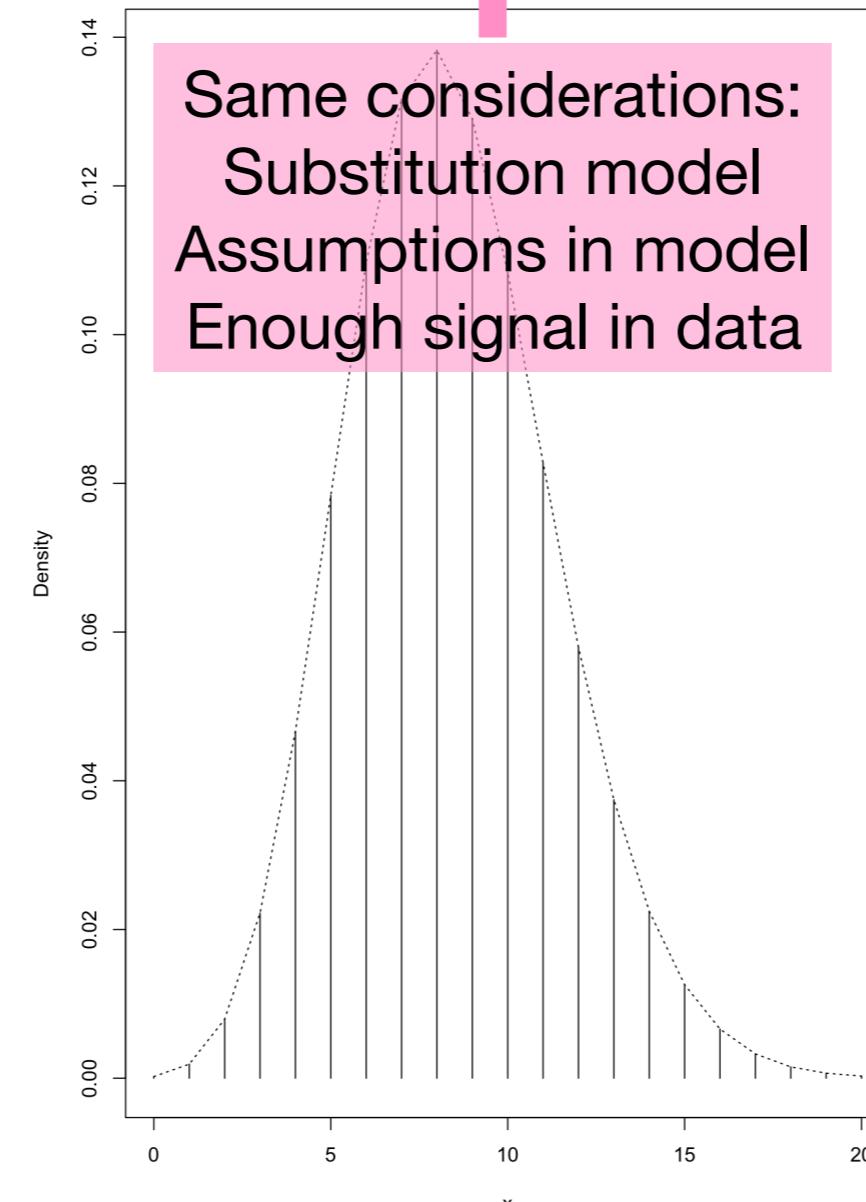
Prior Mean: 0.03

Information in the data

Likelihood

New considerations:  
Which prior model?  
Assumptions in model  
How does this affect inference?

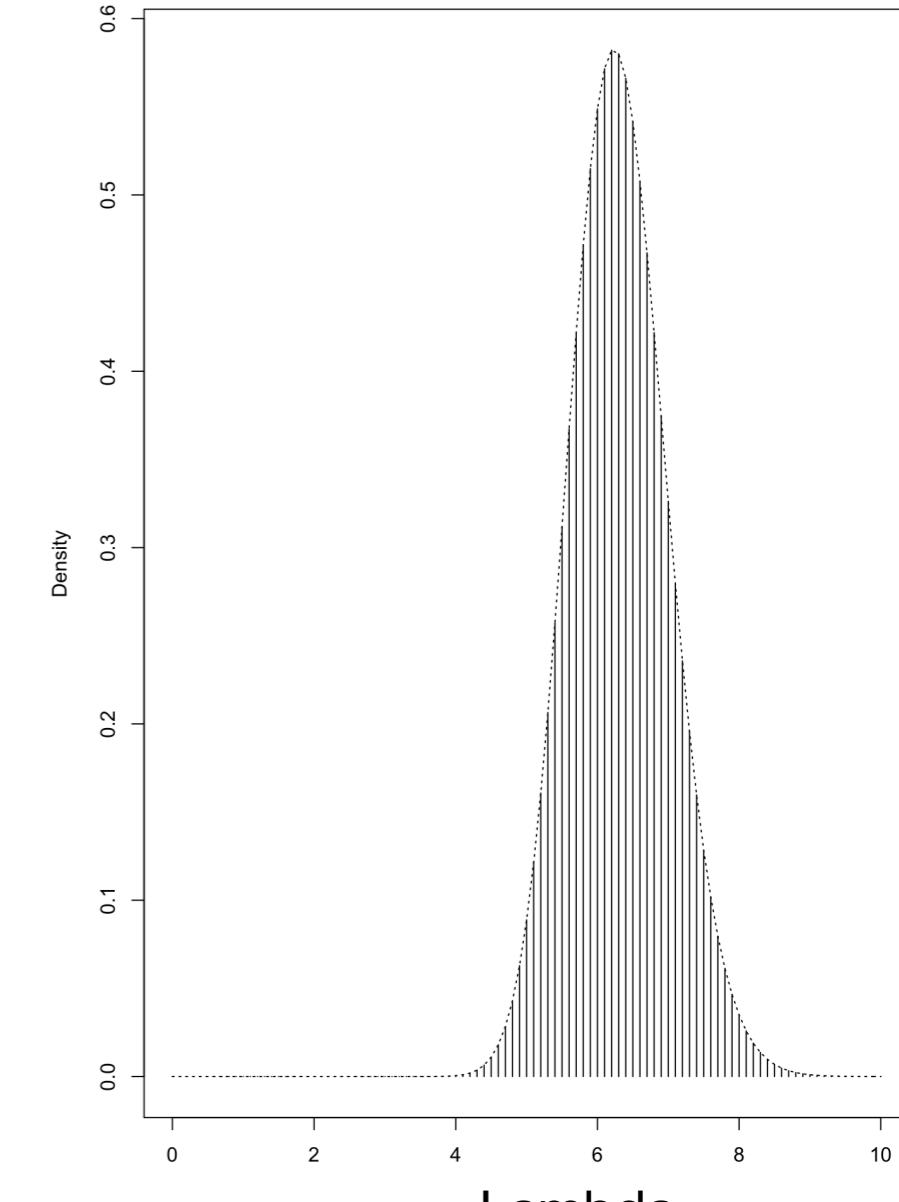
+



Inference on parameter

Posterior

Posterior Mean: 6.31

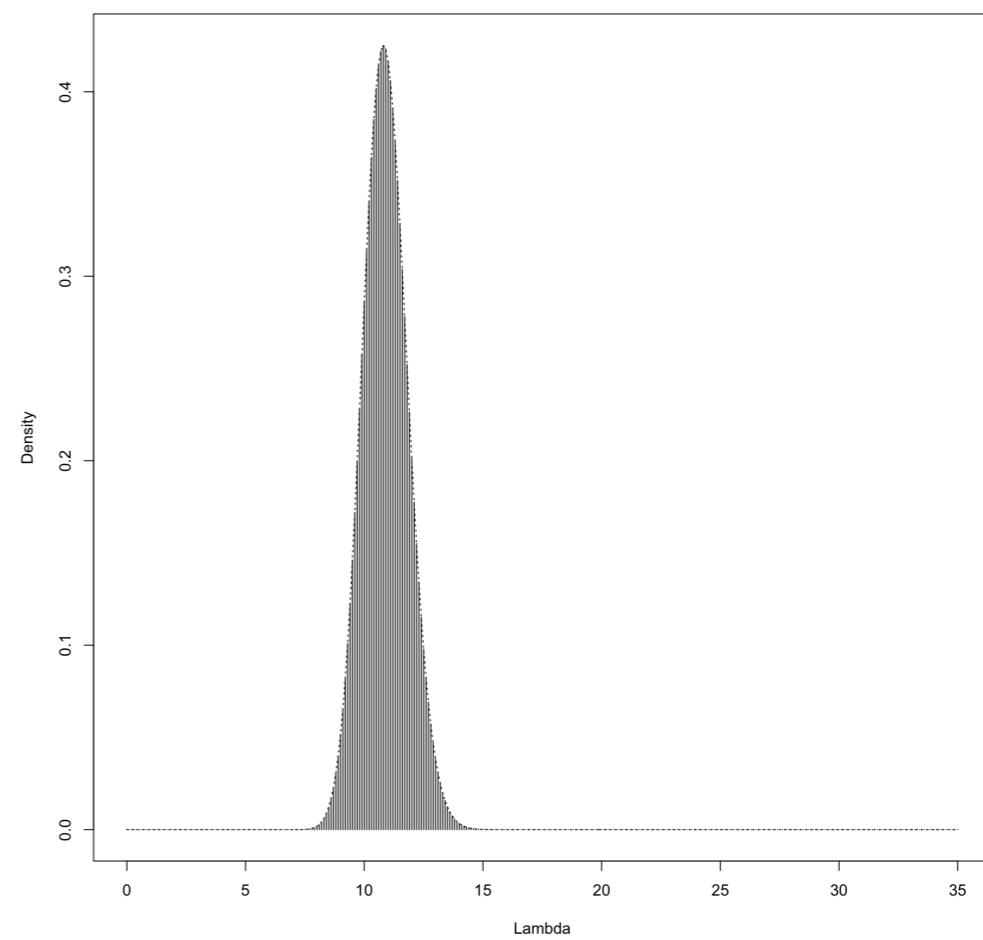
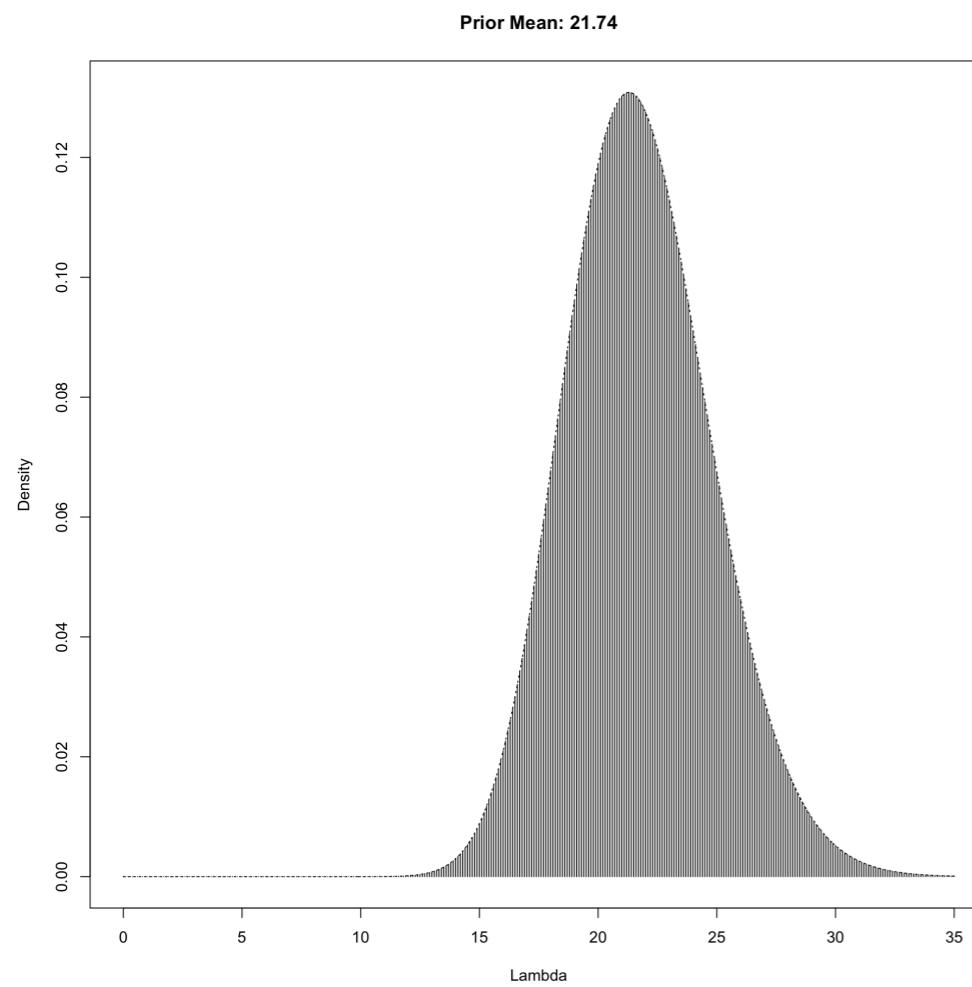
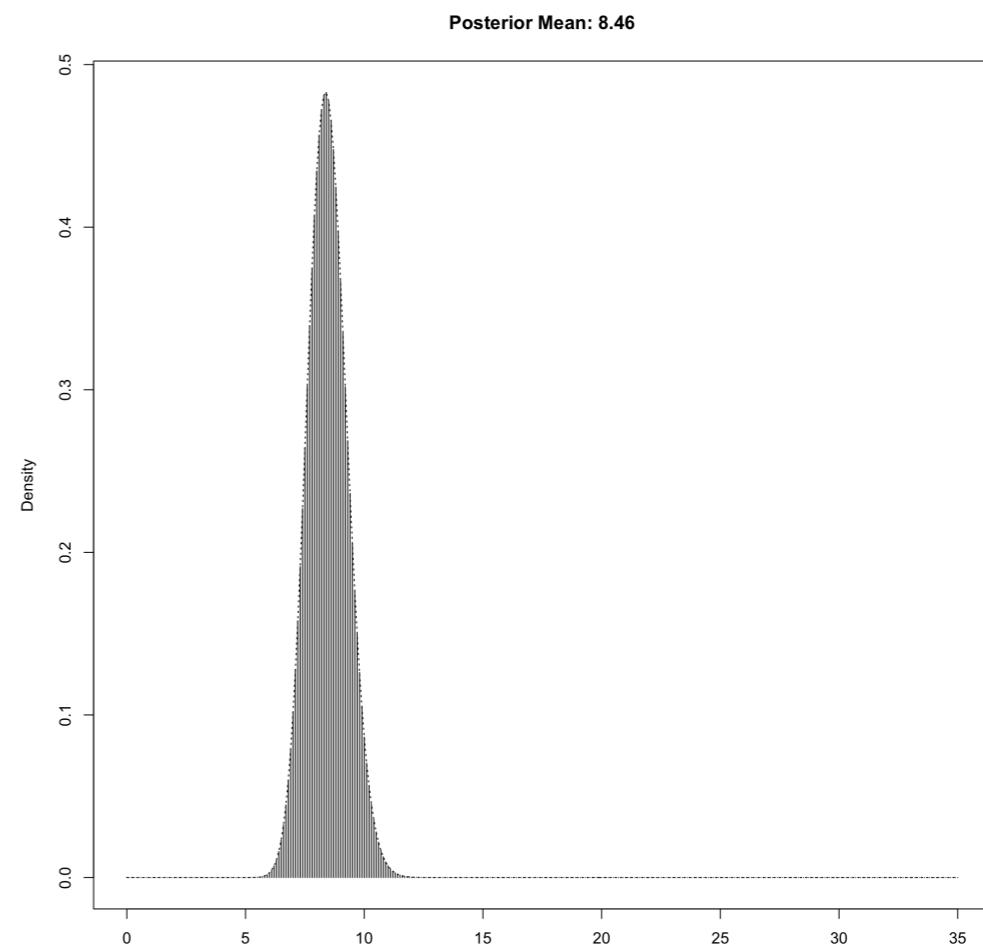
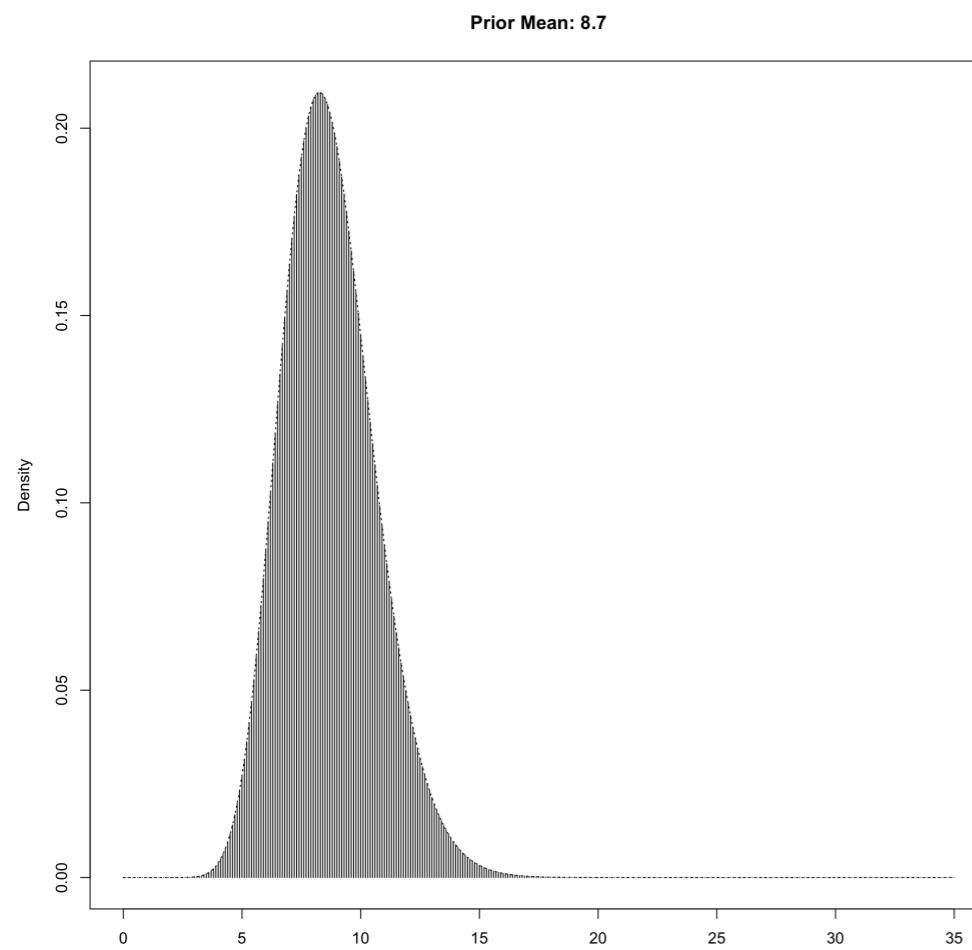


Lambda

$$\lambda \sim \text{Gamma}(\alpha = 0.1, \beta = 3.3)$$

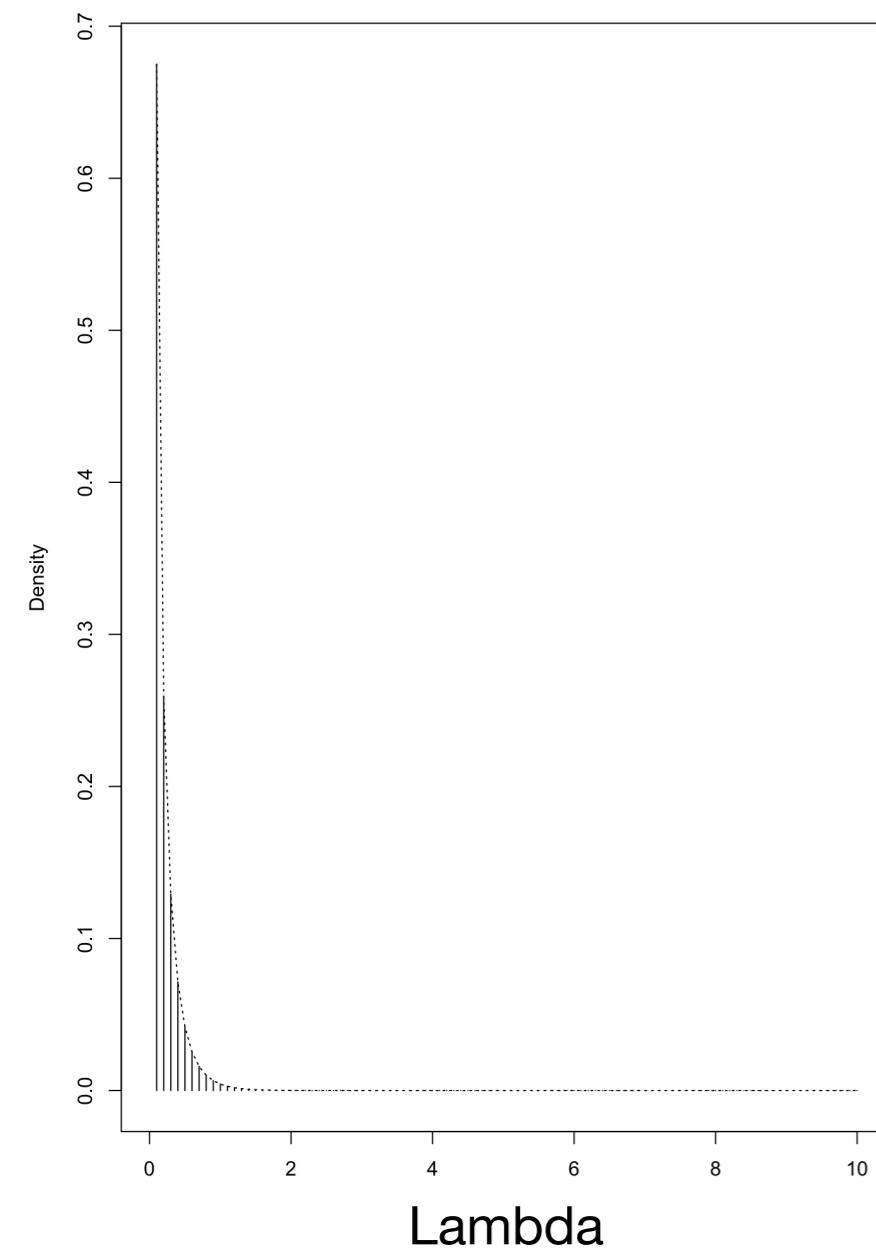
$$\hat{\lambda} = 8.4$$

$$\lambda | \mathbf{X} \sim \text{Gamma}(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$



# Prior

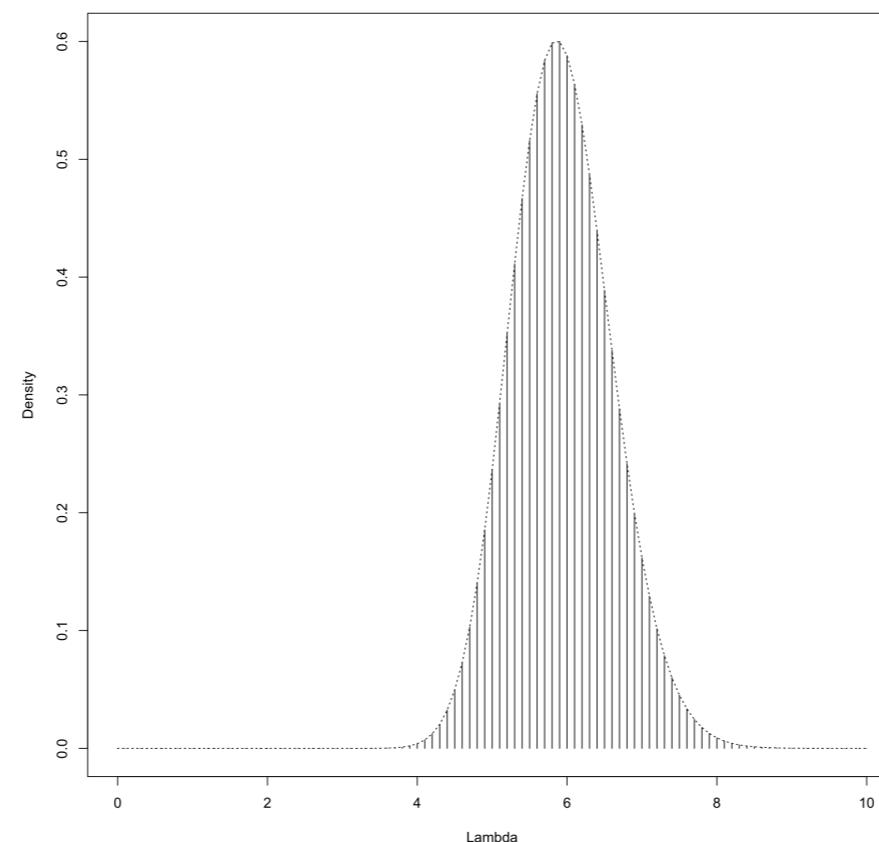
Prior Mean: 0.03



$$\lambda \sim \text{Gamma}(\alpha = 0.1, \beta = 3.3)$$

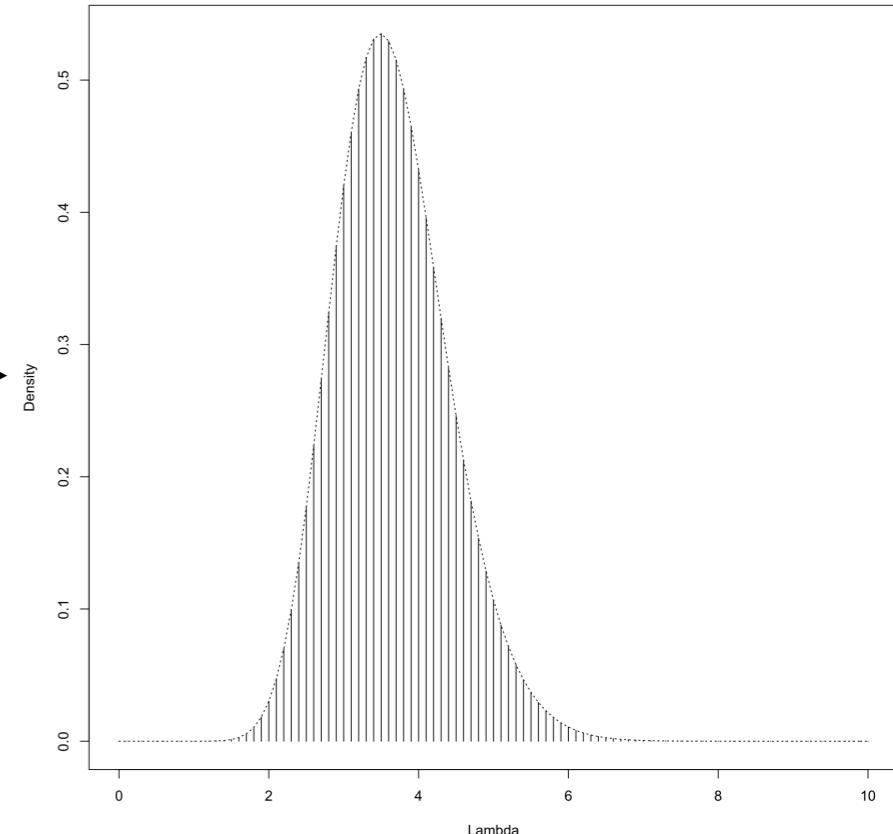
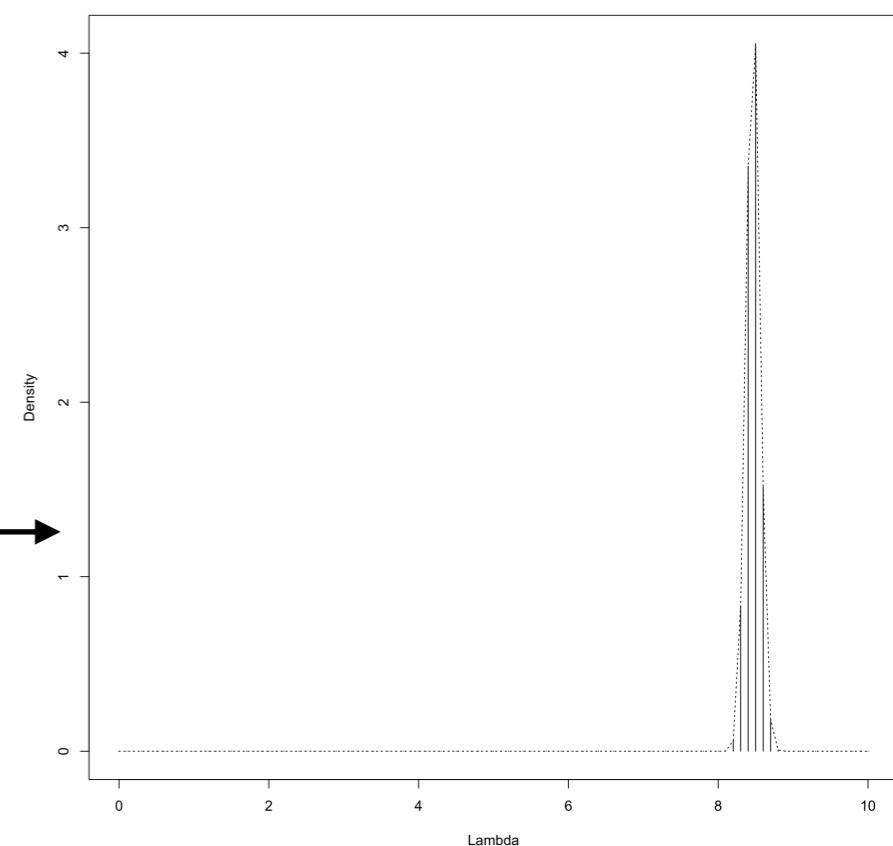
**n=3**

Posterior Mean: 5.93



**n=1000**

Posterior Mean: 8.47



# Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

Your knowledge

Prior

Prior Mean: 0.03

Information in the data

Likelihood

New considerations:  
Which prior model?  
Assumptions in model  
How does this affect inference?

- Pulls the posterior towards it
- With sufficient enough data, it does not matter
- With non-informative/flat priors, you get the same as MLE

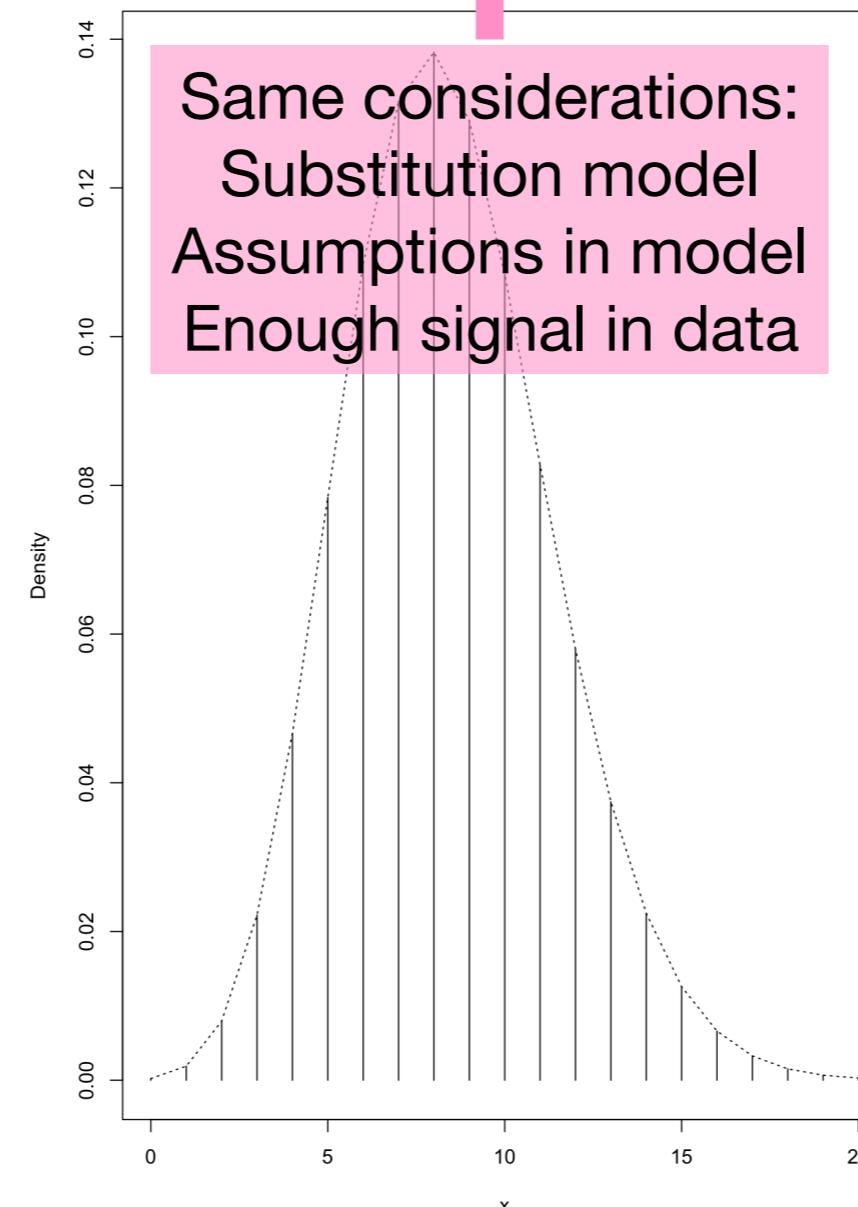
Lambda

$$\lambda \sim \text{Gamma}(\alpha = 0.1, \beta = 3.3)$$

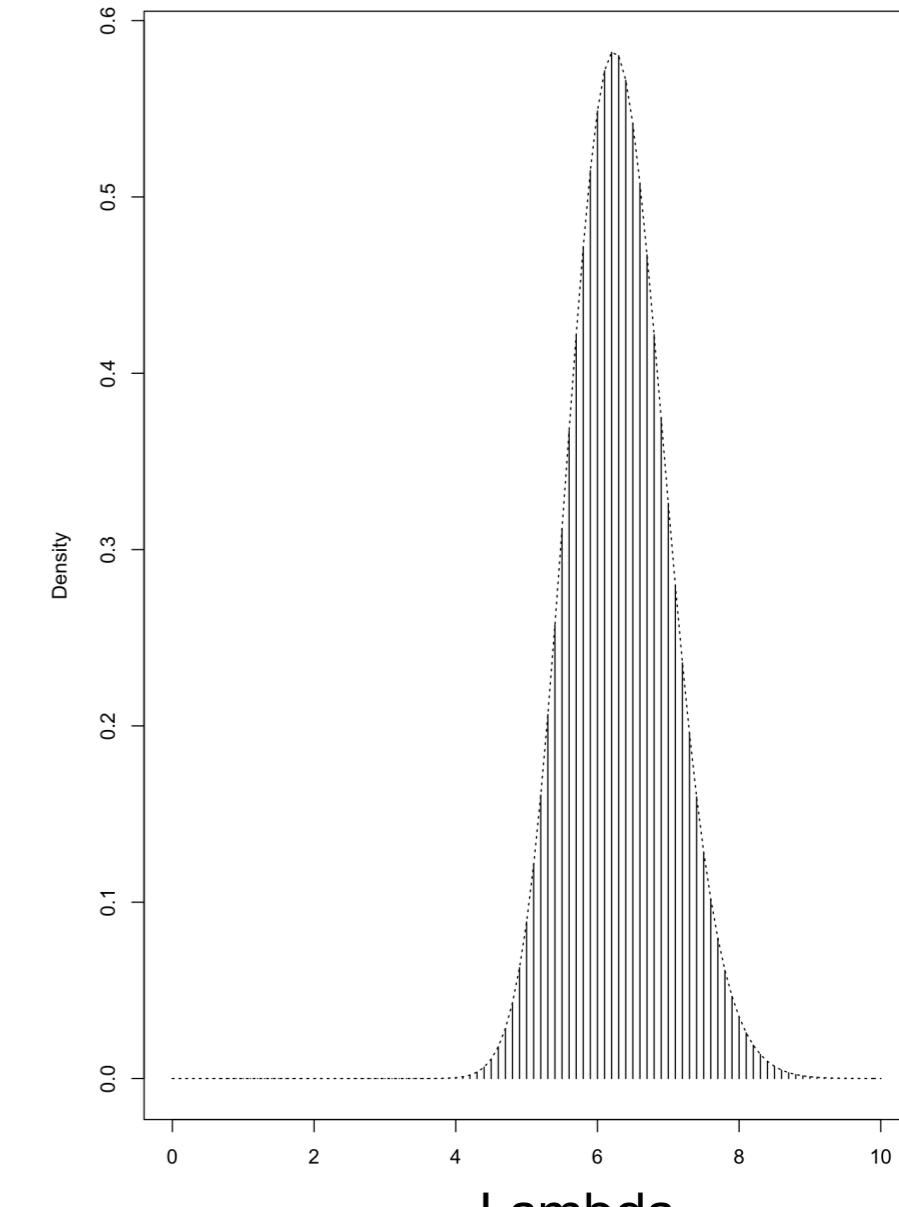
Inference on parameter

Posterior

Posterior Mean: 6.31



$$\hat{\lambda} = 8.4$$



$$\lambda | \mathbf{X} \sim \text{Gamma}(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$

# Posterior distribution

$$P(\lambda|\mathbf{X}) = \frac{L_{\mathbf{X}}(\lambda)p(\lambda)}{p(\mathbf{X})} = \frac{\underset{\text{Likelihood}}{\prod P(X = x_i|\lambda)}}{\underset{\text{Marginal}}{p(\mathbf{X})}} \underset{\text{Prior}}{p(\lambda)}$$

← Many times,  
intractable

# Posterior distribution

$$P(\lambda|\mathbf{X}) = \frac{L_{\mathbf{X}}(\lambda)p(\lambda)}{p(\mathbf{X})} = \frac{\underset{\text{Likelihood}}{\prod P(X = x_i|\lambda)}}{\underset{\text{Marginal}}{p(\mathbf{X})}} \underset{\text{Prior}}{p(\lambda)}$$

← Many times,  
intractable

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

# Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

# Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \left( \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!} \right) \lambda^{\alpha-1} e^{-\beta\lambda}$$

# Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \left( \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!} \right) \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \lambda^{\sum x_i + \alpha - 1} e^{-(n + \beta)\lambda}$$

# Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \left( \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!} \right) \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \lambda^{\sum x_i + \alpha - 1} e^{-(n+\beta)\lambda} \quad \text{← } Gamma\left(\sum x_i + \alpha, n + \beta\right)$$

# Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \left( \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!} \right) \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \lambda^{\sum x_i + \alpha - 1} e^{-(n+\beta)\lambda} \quad \text{← } Gamma\left(\sum x_i + \alpha, n + \beta\right)$$

$$P(\lambda|\mathbf{X}) = \frac{(n+\beta)^{(\sum x_i + \alpha)}}{\Gamma(\sum x_i + \alpha)} \lambda^{\sum x_i + \alpha - 1} e^{-(n+\beta)\lambda}$$

# Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$P(\lambda|\mathbf{X}) = \frac{(n+\beta)^{(\sum x_i + \alpha)}}{\Gamma(\sum x_i + \alpha)} \lambda^{\sum x_i + \alpha - 1} e^{-(n+\beta)\lambda}$$

$$\lambda|\mathbf{X} \sim Gamma\left(\sum x_i + \alpha, n + \beta\right)$$

# Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta) \quad \text{Conjugate prior}$$

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$P(\lambda|\mathbf{X}) = \frac{(n+\beta)^{(\sum x_i + \alpha)}}{\Gamma(\sum x_i + \alpha)} \lambda^{\sum x_i + \alpha - 1} e^{-(n+\beta)\lambda}$$

$$\lambda|\mathbf{X} \sim Gamma \left( \sum x_i + \alpha, n + \beta \right)$$

# Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta) \quad \text{Conjugate prior}$$

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$P(\lambda|\mathbf{X}) = \frac{(n+\beta)^{(\sum x_i + \alpha)}}{\Gamma(\sum x_i + \alpha)} \lambda^{\sum x_i + \alpha - 1} e^{-(n+\beta)\lambda}$$

$$\lambda|\mathbf{X} \sim Gamma\left(\sum x_i + \alpha, n + \beta\right)$$

Posterior mean:  $\frac{\sum x_i + \alpha}{n + \beta} = \frac{n}{n + \beta} \left( \frac{\sum x_i}{n} \right) + \frac{\beta}{n + \beta} \left( \frac{\alpha}{\beta} \right)$

# Intractable posterior distribution

$$X_i | \lambda \sim Poisson(\lambda)$$

$$\lambda \sim Lognormal(\mu, \sigma)$$

$$P(\lambda | \mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{1}{\lambda \sigma \sqrt{2\pi}} \exp \left( -\frac{(\log \lambda - \mu)^2}{2\sigma^2} \right)$$

# Intractable posterior distribution

$$X_i | \lambda \sim Poisson(\lambda)$$

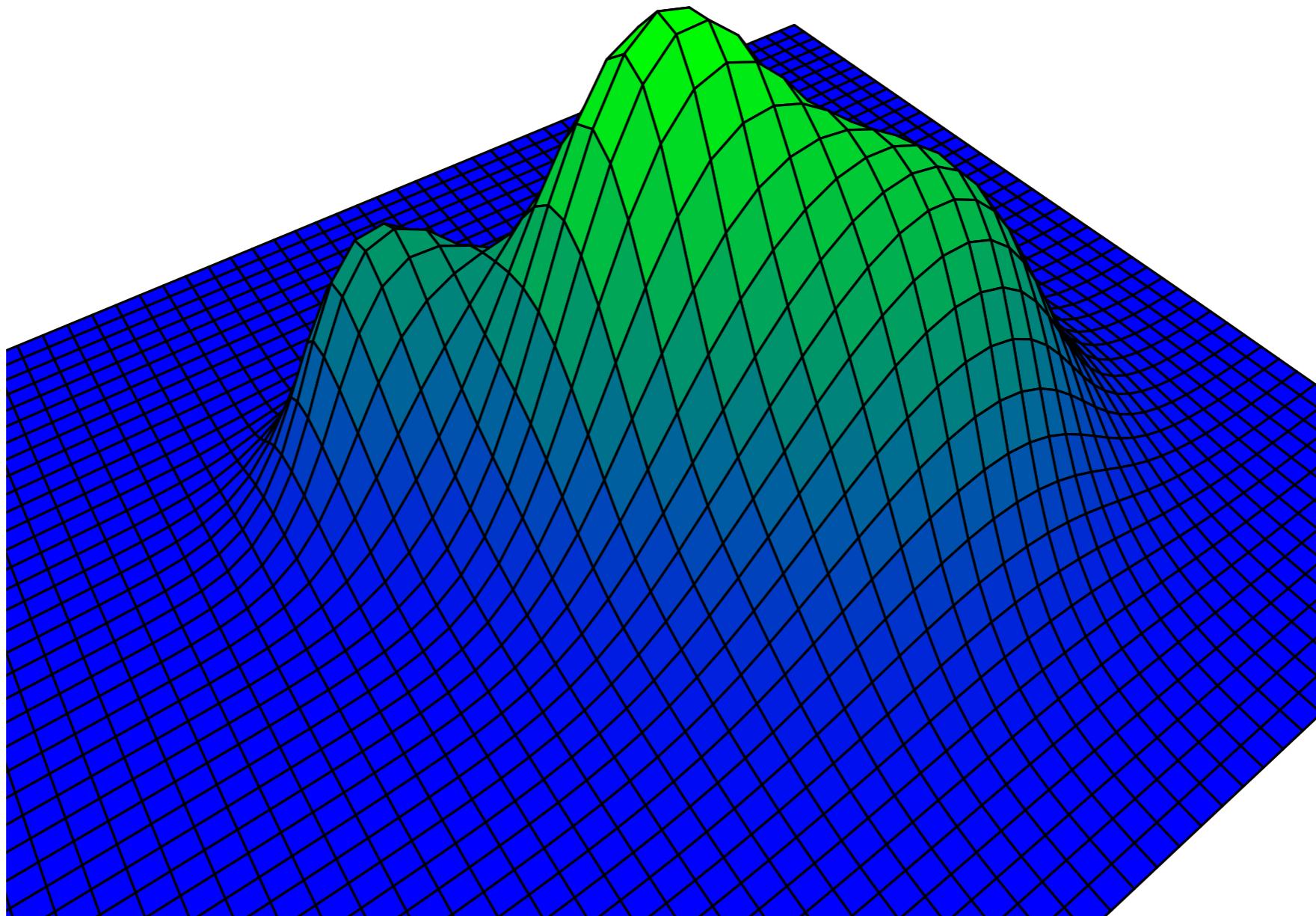
$$\lambda \sim Lognormal(\mu, \sigma)$$

$$P(\lambda | \mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{1}{\lambda \sigma \sqrt{2\pi}} \exp \left( -\frac{(\log \lambda - \mu)^2}{2\sigma^2} \right)$$



**Not a known  
distribution anymore  
We need to  
approximate it**

# MCMC: A way to approximate intractable posterior distributions



# MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{1}{\lambda \sigma \sqrt{2\pi}} \exp\left(-\frac{(\log \lambda - \mu)^2}{2\sigma^2}\right)$$

**Initialization:** Start at a random  $\lambda_0$

**Loop:** For  $i$ , propose a new lambda

$$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If  $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$

- Else, accept  $\lambda_i = \lambda^*$  with probability
- Otherwise,  $\lambda_i = \lambda_{i-1}$

$$\alpha = \frac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$$

# MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{1}{\lambda \sigma \sqrt{2\pi}} \exp\left(-\frac{(\log \lambda - \mu)^2}{2\sigma^2}\right)$$

**Initialization:** Start at a random  $\lambda_0$

**Loop:** For  $i$ , propose a new lambda

$$\lambda^* \sim \text{Uniform}(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If  $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$

$$\alpha = \frac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$$

- Else, accept  $\lambda_i = \lambda^*$  with probability  $\alpha$
- Otherwise,  $\lambda_i = \lambda_{i-1}$

**Proposal distribution**

Usually symmetric,  
but could be  
asymmetric  
(Hasting ratio)

# MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{1}{\lambda \sigma \sqrt{2\pi}} \exp\left(-\frac{(\log \lambda - \mu)^2}{2\sigma^2}\right)$$

**Initialization:** Start at a random  $\lambda_0$

**Loop:** For  $i$ , propose a new lambda

$$\lambda^* \sim \text{Uniform}(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If  $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$

$$\alpha = \frac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$$

- Else, accept  $\lambda_i = \lambda^*$  with probability  $\alpha$
- Otherwise,  $\lambda_i = \lambda_{i-1}$

**Proposal distribution**

Usually symmetric,  
but could be  
asymmetric  
(Hasting ratio)

**Acceptance probability**

Posterior ratio

# MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{1}{\lambda \sigma \sqrt{2\pi}} \exp\left(-\frac{(\log \lambda - \mu)^2}{2\sigma^2}\right)$$

**Initialization:** Start at a random  $\lambda_0$

**Loop:** For  $i$ , propose a new lambda

$$\lambda^* \sim \text{Uniform}(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If  $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept  $\lambda_i = \lambda^*$  with probability
- Otherwise,  $\lambda_i = \lambda_{i-1}$

$$\alpha = \frac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$$

**Proposal distribution**

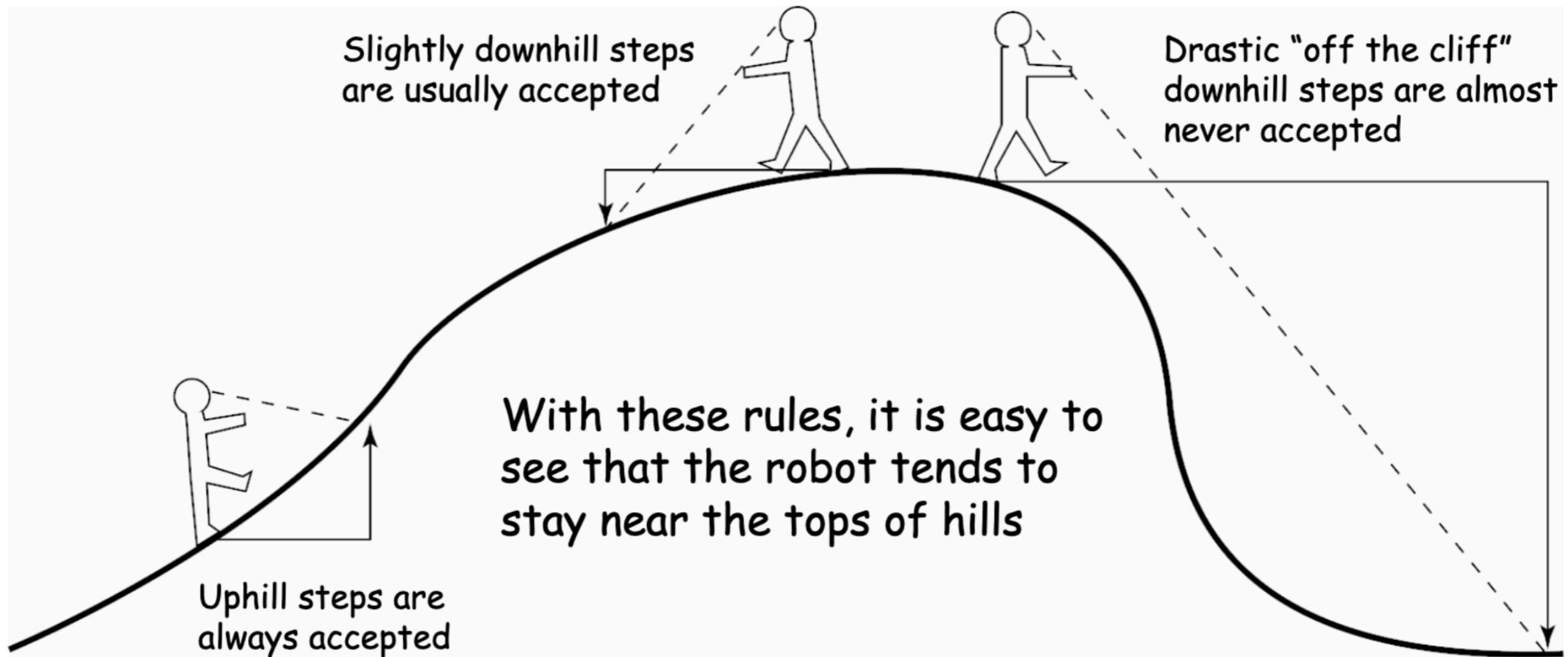
Usually symmetric,  
but could be  
asymmetric  
(Hasting ratio)

**Acceptance probability**

Posterior ratio

The chain tends to stay on regions of high posterior

# MCMC robot's rules



# Hastings ratio

$$\alpha = \left[ \frac{P(\lambda^* | \mathbf{X})}{P(\lambda | \mathbf{X})} \right] \left[ \frac{q(\lambda | \lambda^*)}{q(\lambda^* | \lambda)} \right]$$

Proposal distribution

**Posterior ratio**

**Hastings ratio**

# MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left( \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \frac{1}{\lambda \sigma \sqrt{2\pi}} \exp\left(-\frac{(\log \lambda - \mu)^2}{2\sigma^2}\right)$$

**Initialization:** Start at a random  $\lambda_0$

**Loop:** For  $i$ , propose a new lambda

$$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If  $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$

$$\alpha = \frac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$$

- Else, accept  $\lambda_i = \lambda^*$  with probability

- Otherwise,  $\lambda_i = \lambda_{i-1}$

**Optional (highly recommended) homework:**

- Code the MCMC for this example in your preferred programming language
- Plot the MCMC histogram and compute the posterior mean for lambda
- Play with different values for  $w$  and  $n$  (length of the chain)

# MCMC considerations

- Choice of priors
- Mixing
- Convergence
- Burnin

# MCMC considerations

- Choice of priors
- Mixing ← How well you navigate the parameter space?
- Convergence
- Burnin

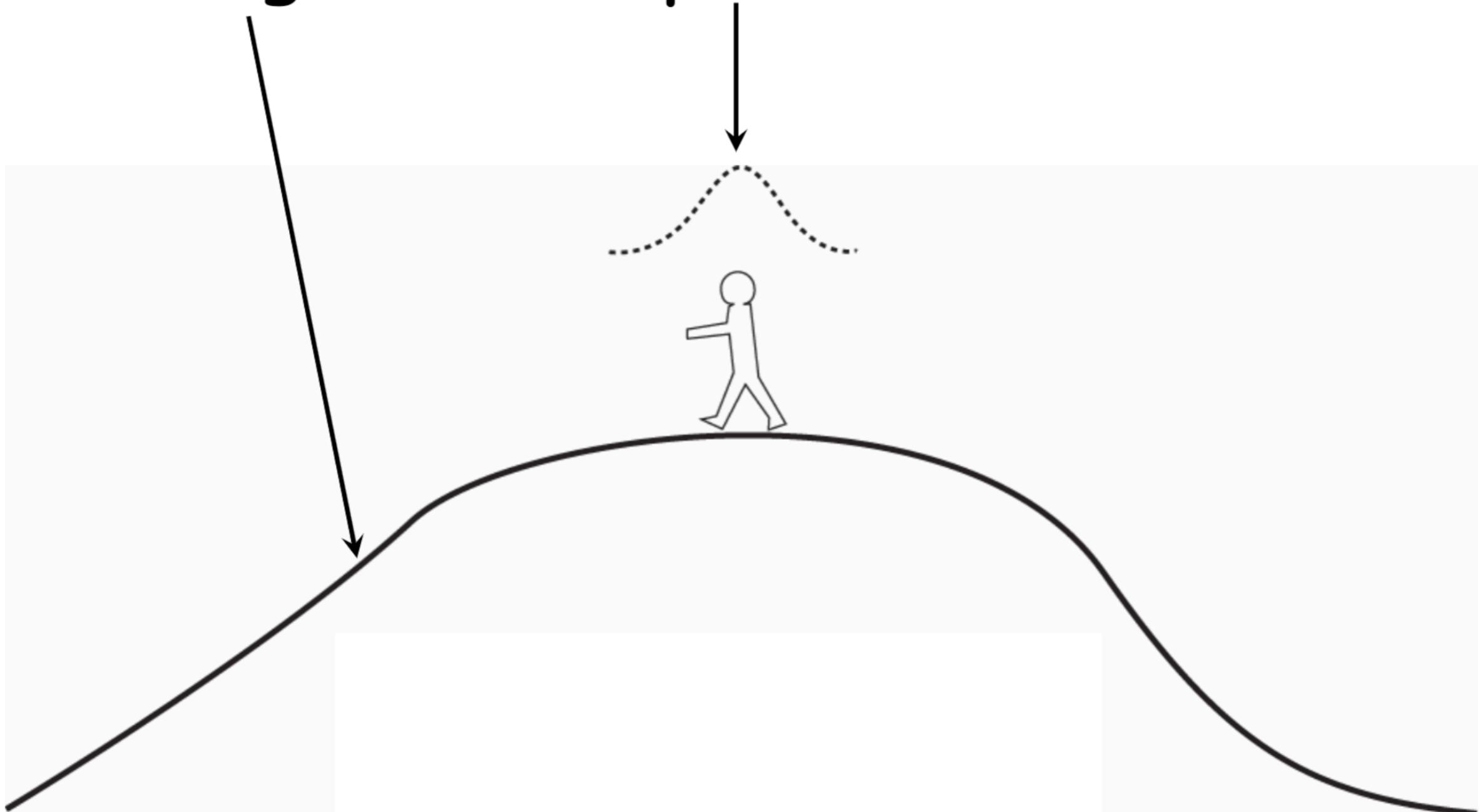
# MCMC considerations

- Choice of priors
- Mixing ← How well you navigate the parameter space?
- Convergence ← How well you reach all regions of high posterior values?
- Burnin

# MCMC considerations

- Choice of priors
- Mixing ← How well you navigate the parameter space?
- Convergence ← How well you reach all regions of high posterior values?
- Burnin ← How long it takes to reach regions of high posterior values?

# Target vs. Proposal Distributions



# Target vs. Proposal Distributions

Proposal distributions  
with **smaller variance**...



**Disadvantage:** robot takes  
smaller steps, more time  
required to explore the  
same area

**Advantage:** robot seldom  
refuses to take proposed  
steps

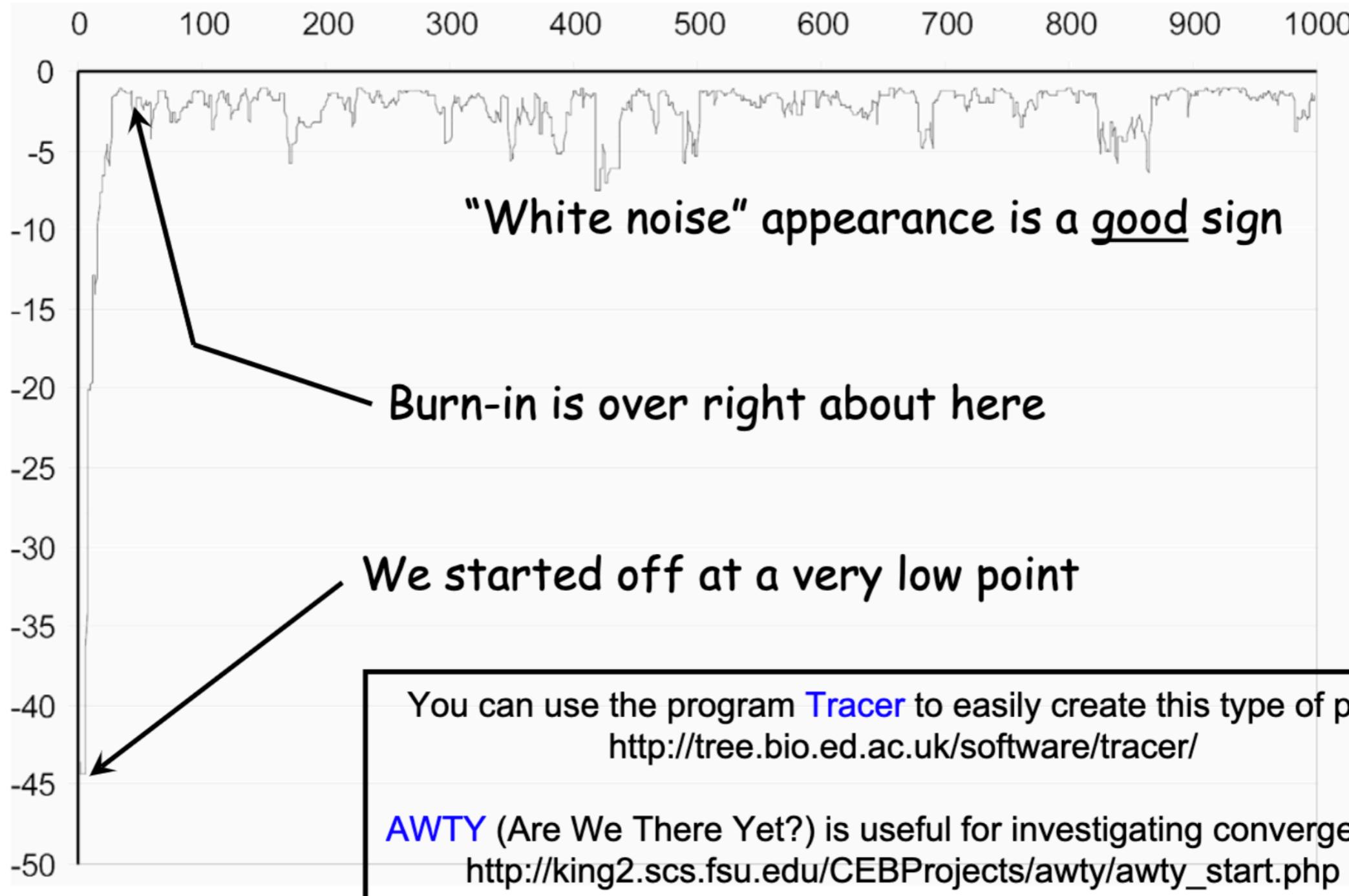
# Target vs. Proposal Distributions

Proposal distributions  
with **larger variance**...

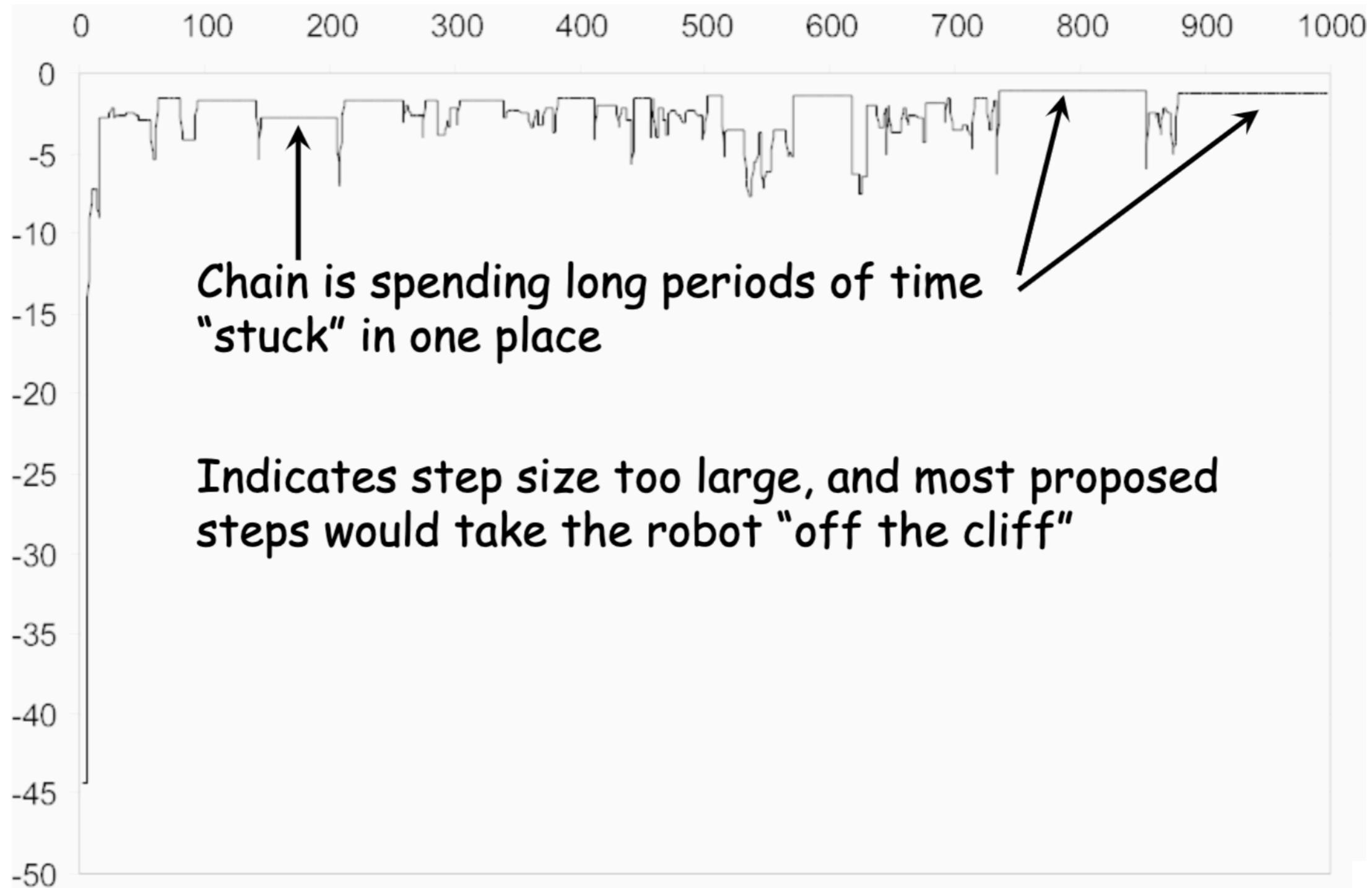
**Disadvantage:** robot  
often proposes a step  
that would take it off  
a cliff, and refuses to  
move



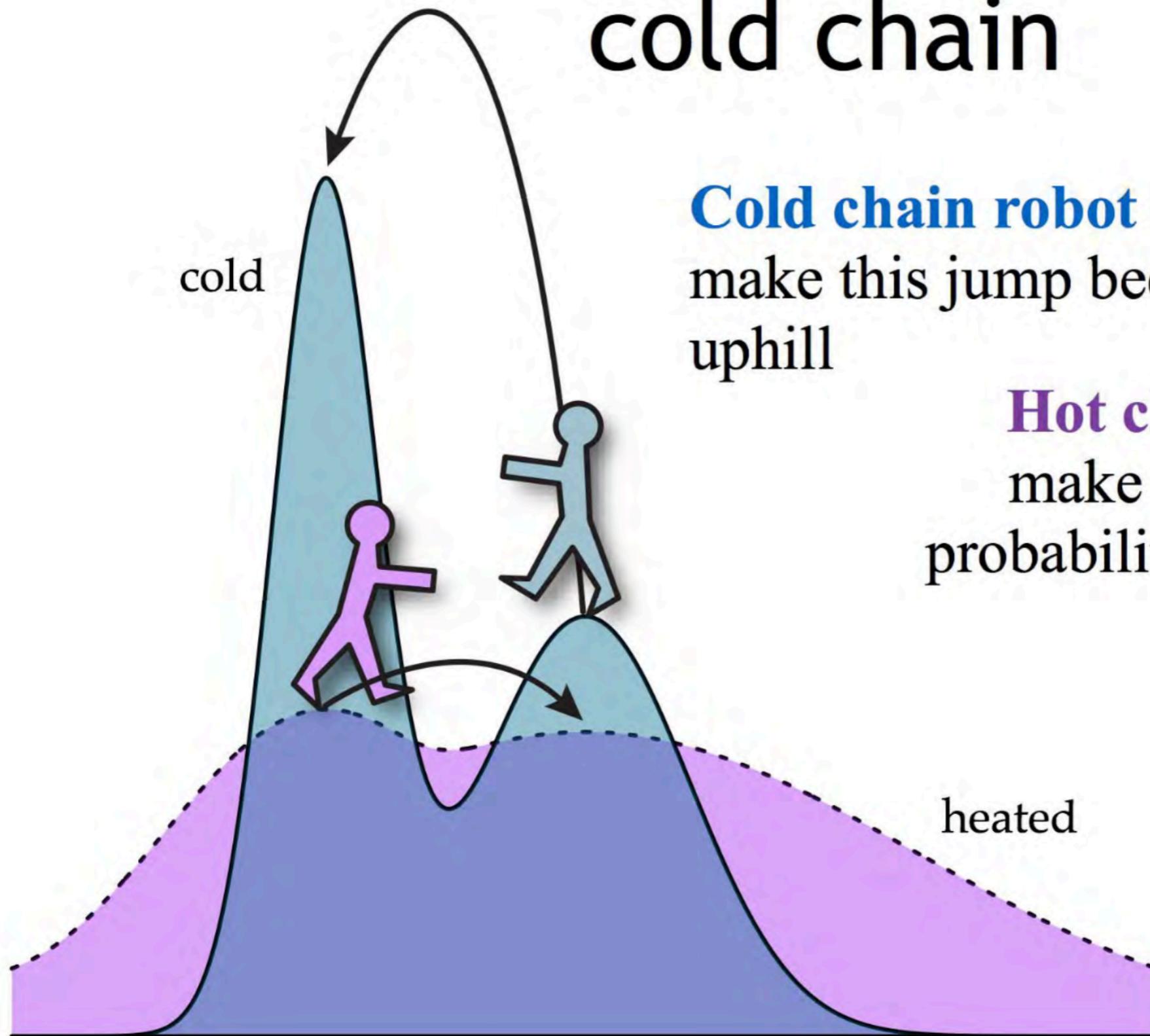
# Trace plots



# Poor mixing



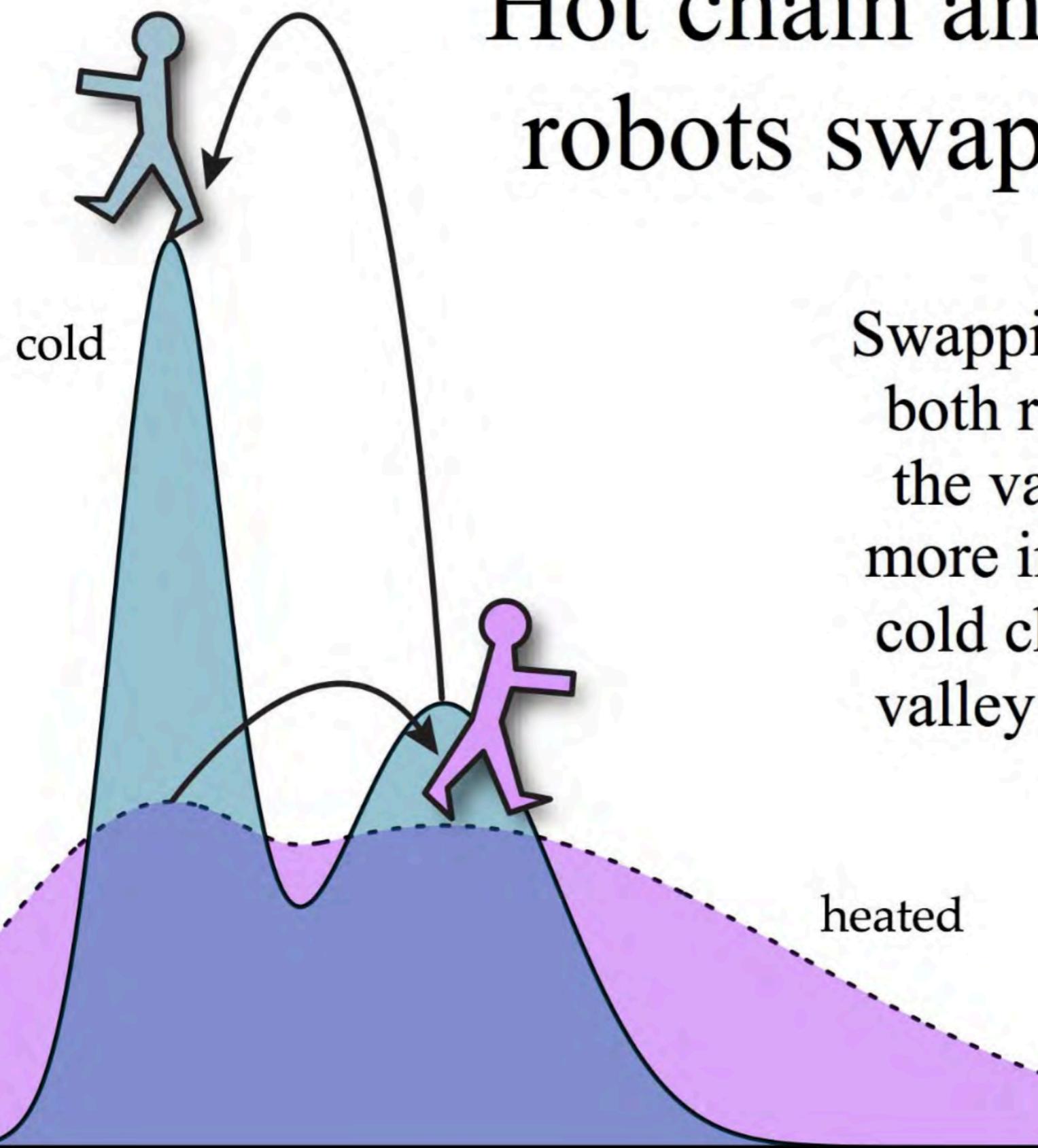
# Heated chains act as scouts for the cold chain



**Cold chain robot** can easily make this jump because it is uphill

**Hot chain robot** can also make this jump with high probability because it is only slightly downhill

# Hot chain and cold chain robots swapping places



Swapping places means both robots can cross the valley, but this is more important for the cold chain because its valley is much deeper

“Metropolis algorithm will produce a precise and accurate approximation of the posterior distribution if run long enough”. - Paul Lewis

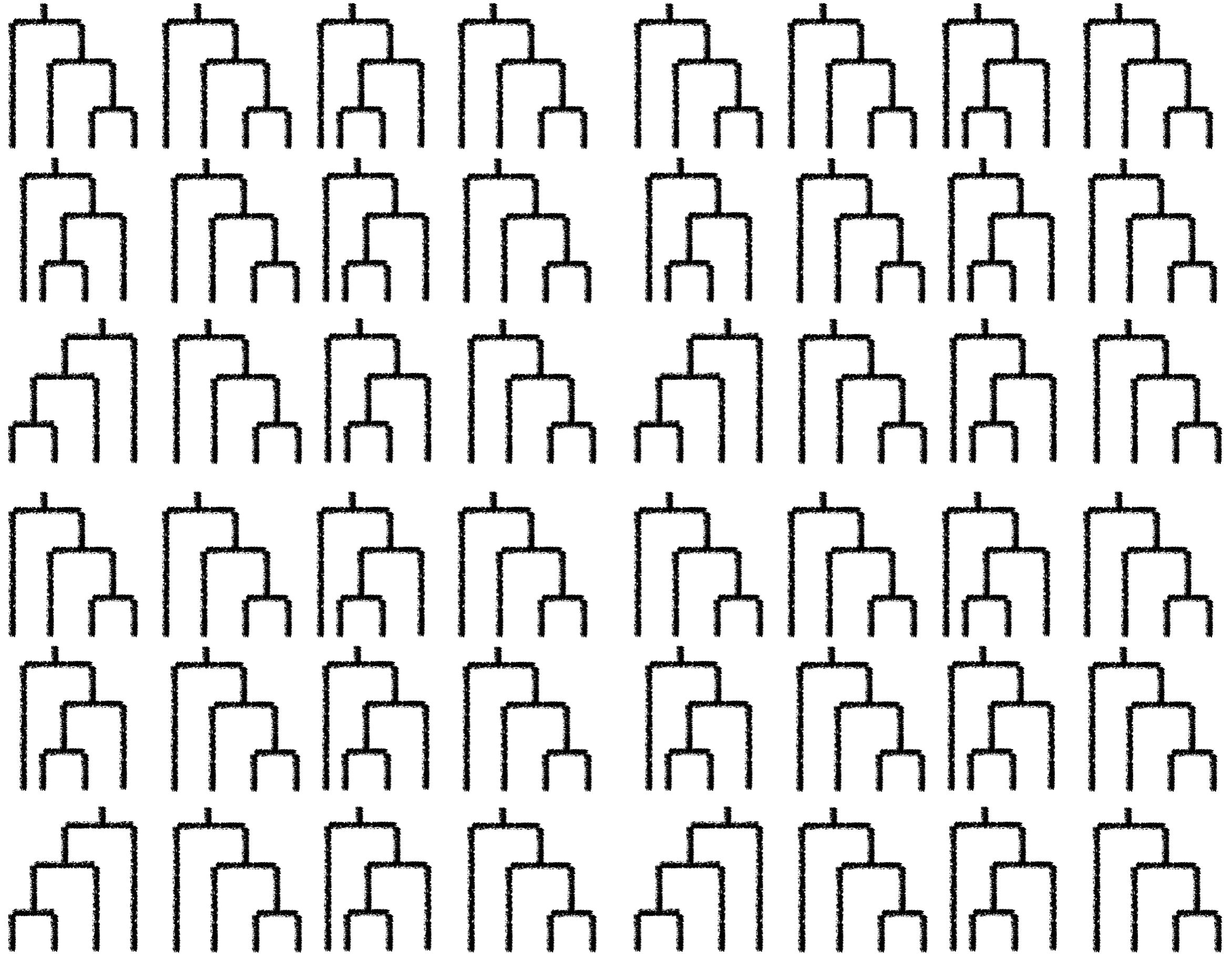
“People always forget how long of a time infinity really is” - paraphrased from Dave Swofford

**Back to  
phylogenetics**

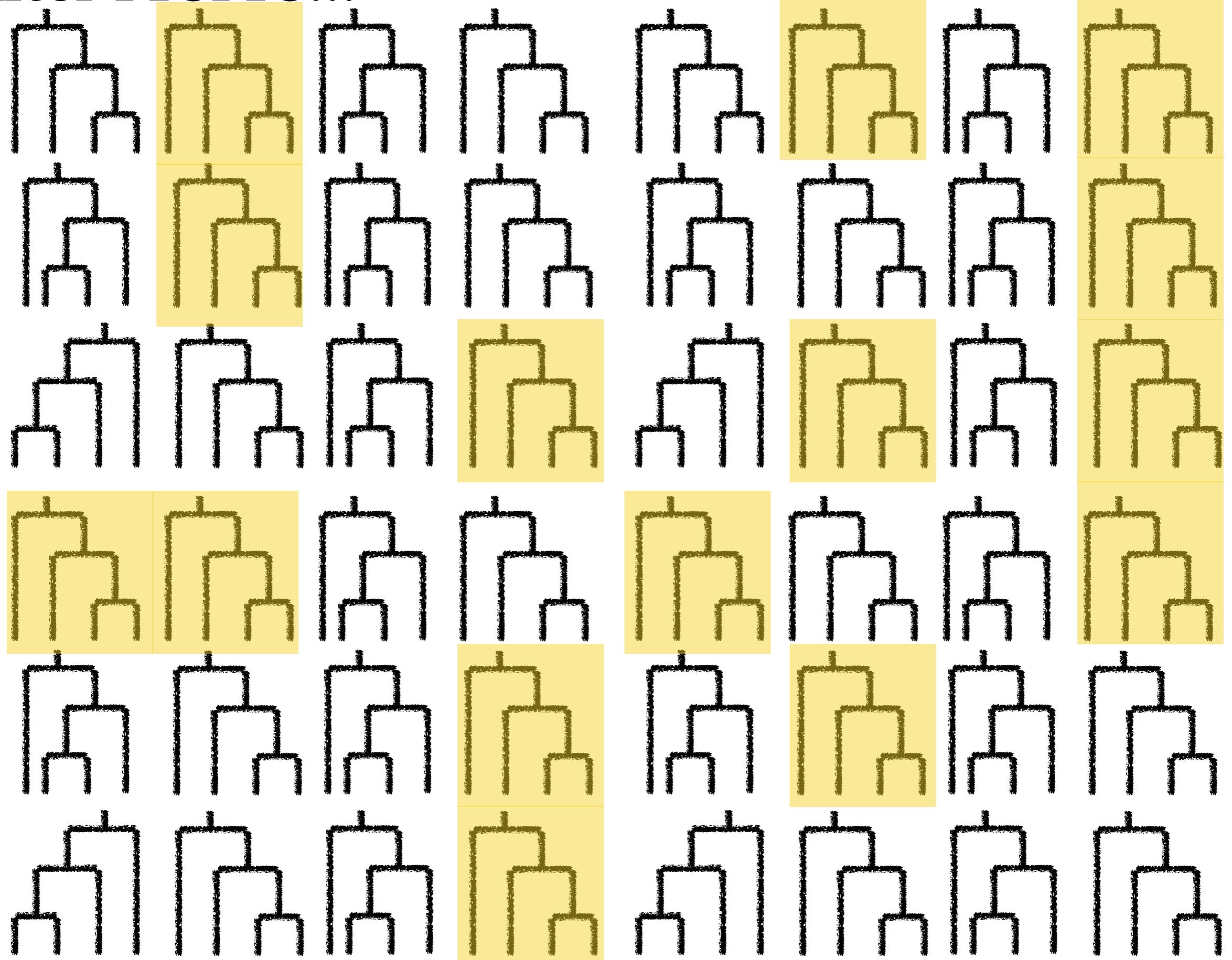
- **Start with** random tree and arbitrary initial values for branch lengths and model parameters
- **Each generation** consists of one of these (chosen at random):
  - Propose a **new tree** (e.g. Larget-Simon move) and either accept or reject the move
  - Propose (and either accept or reject) a **new model parameter value**
- Every  $k$  generations, save tree topology, branch lengths and all model parameters (i.e. **sample the chain**)
- After  $n$  generations, **summarize sample** using histograms, means, credible intervals, etc.

After MCMC...

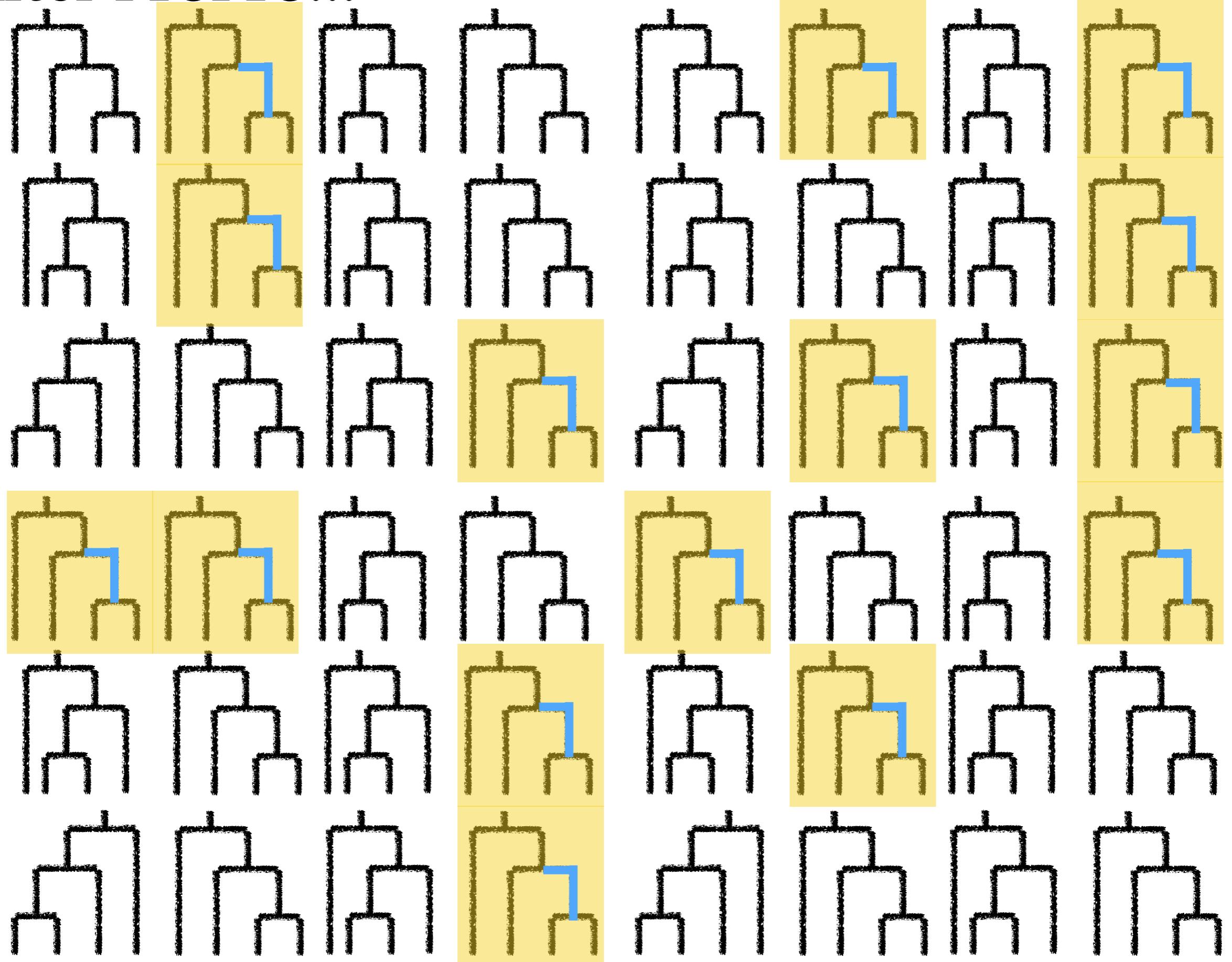
After MCMC...

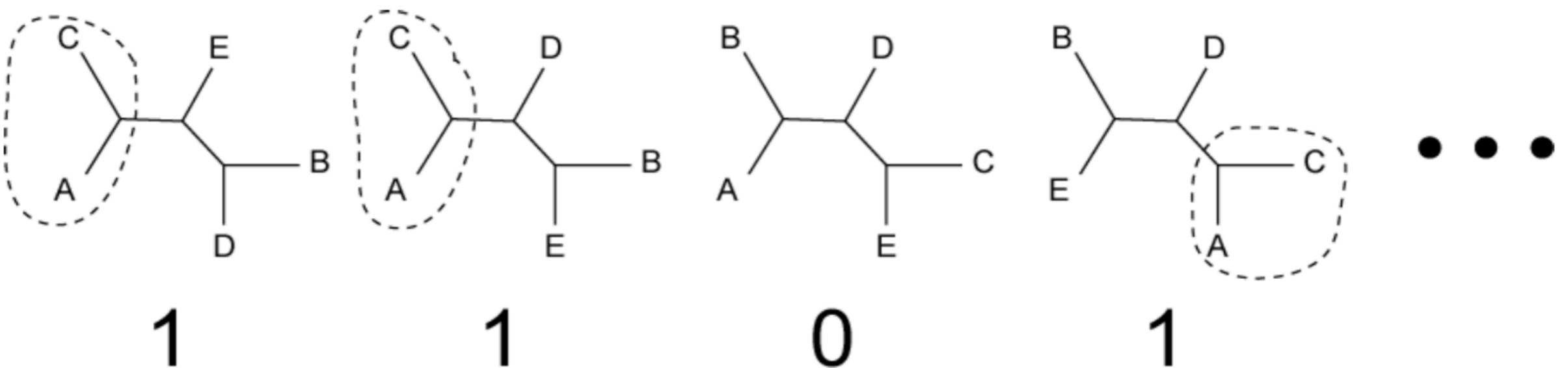


After MCMC...



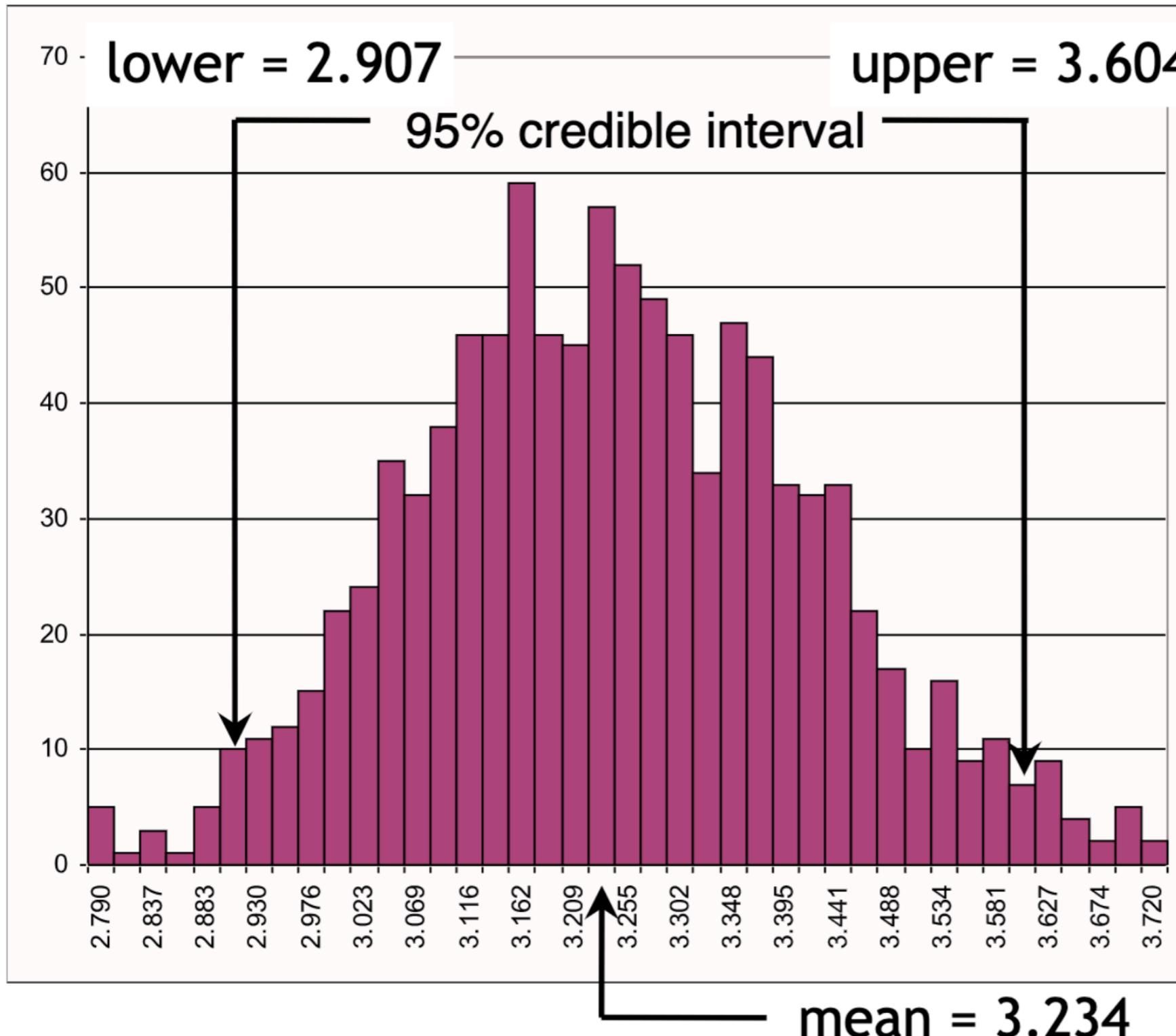
After MCMC...





The posterior probability of the split  $AC \mid BDE$  may be approximated by the fraction of trees sampled from the posterior that contain that split.

# Marginal Posterior Distribution of $\kappa$

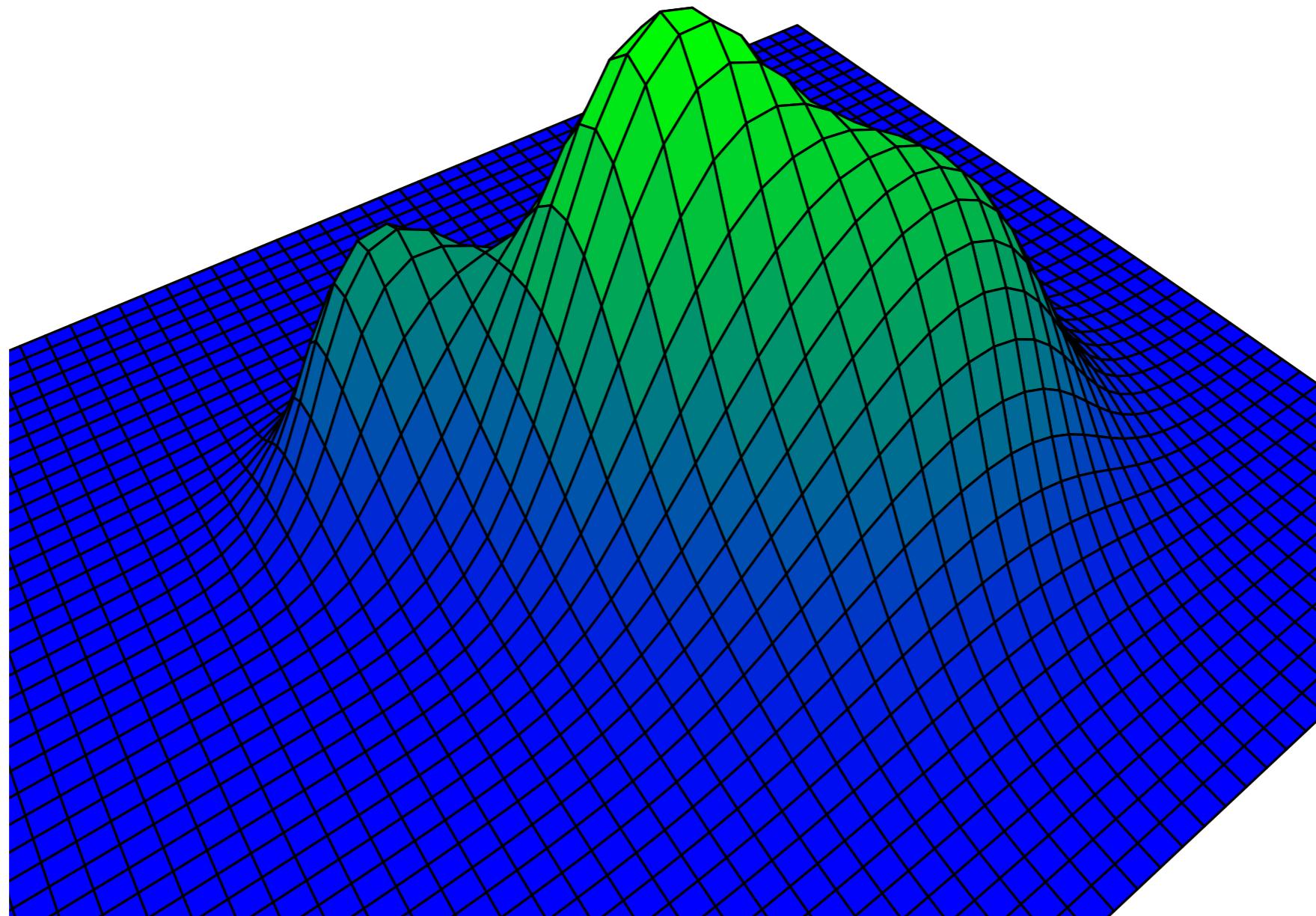


Histogram created from a sample of 1000 kappa values.

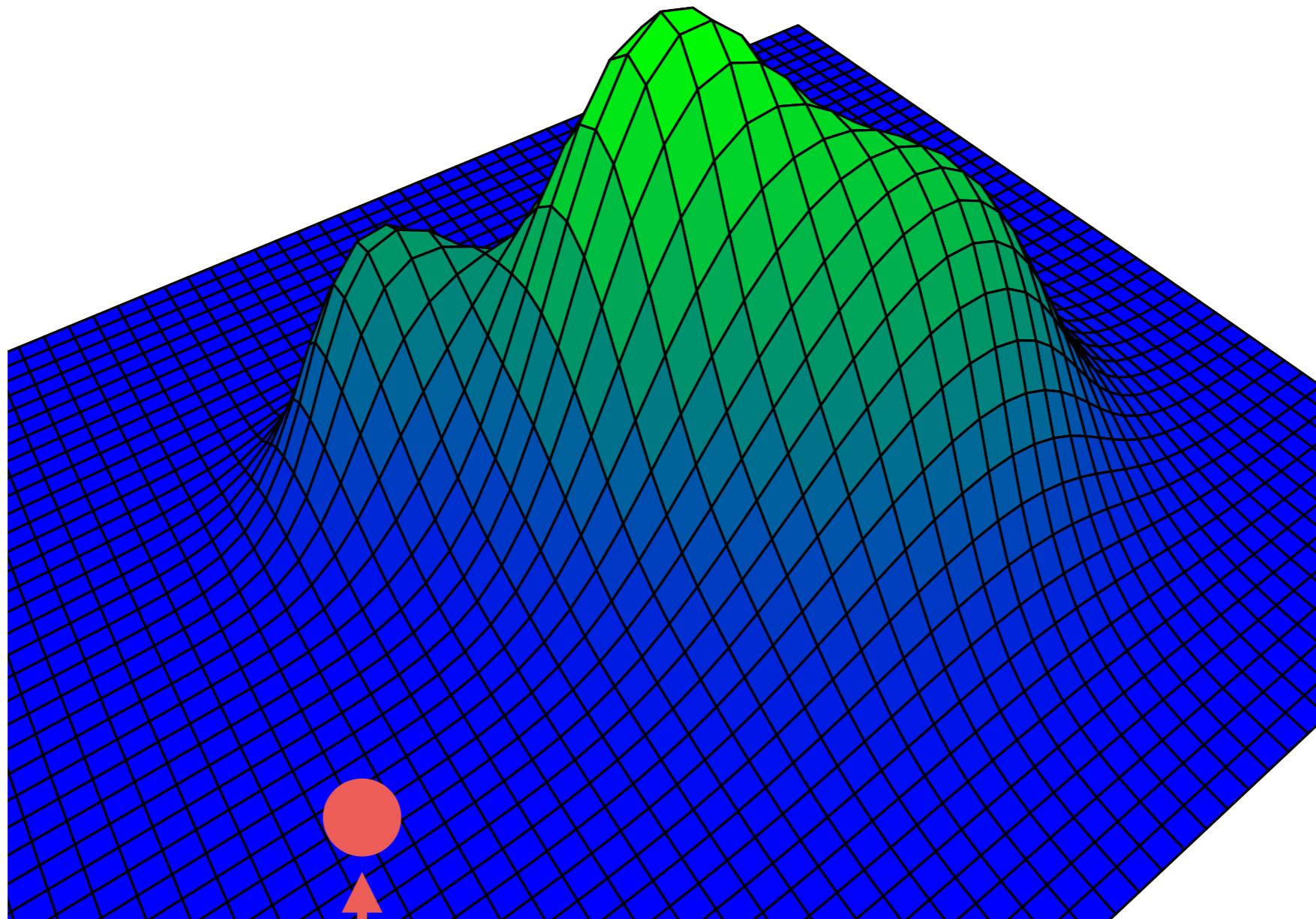
# Why is MCMC so slow?

Why is MCMC so slow? Traverse tree space

Why is MCMC so slow? Traverse tree space

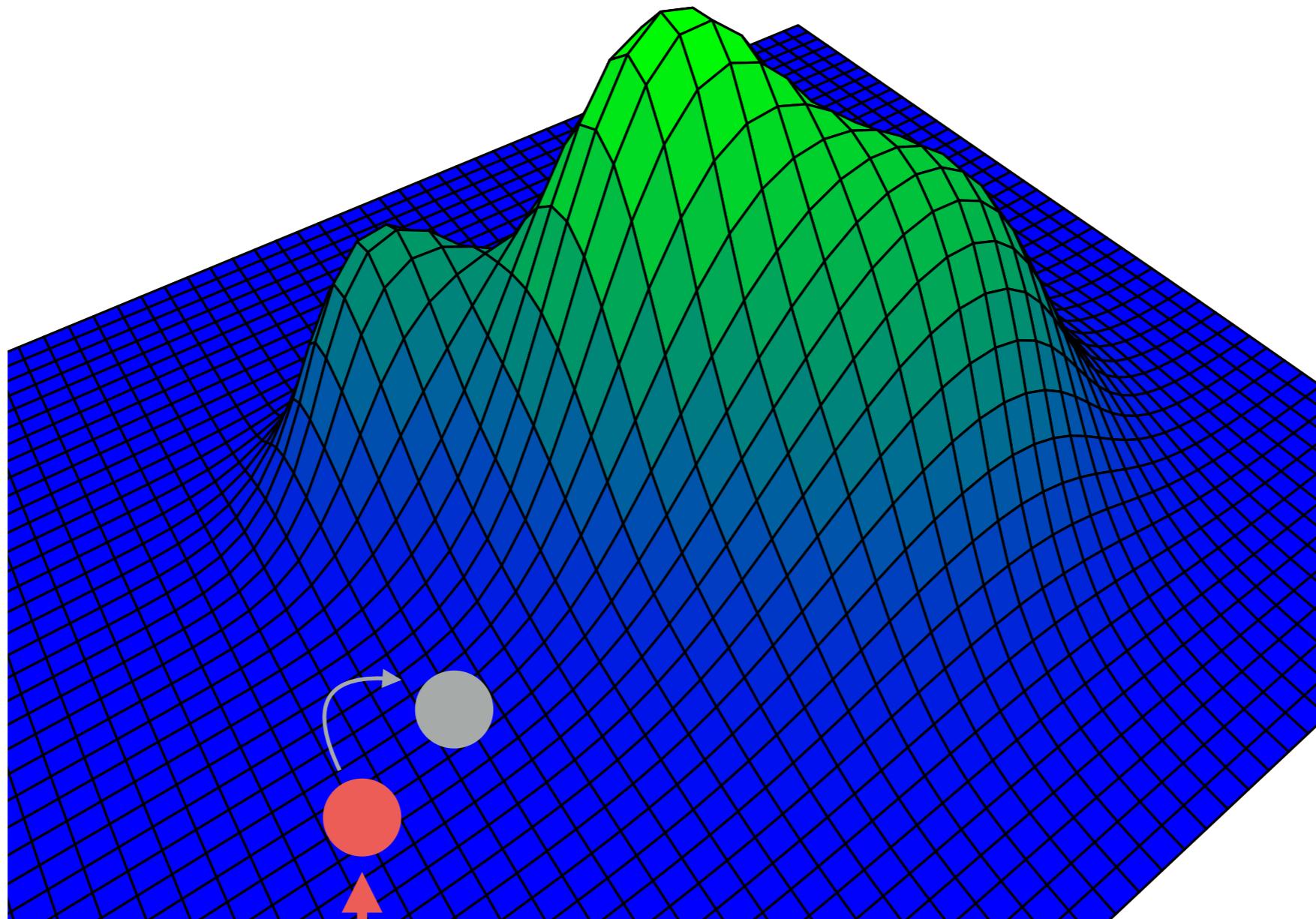


# Why is MCMC so slow? Traverse tree space



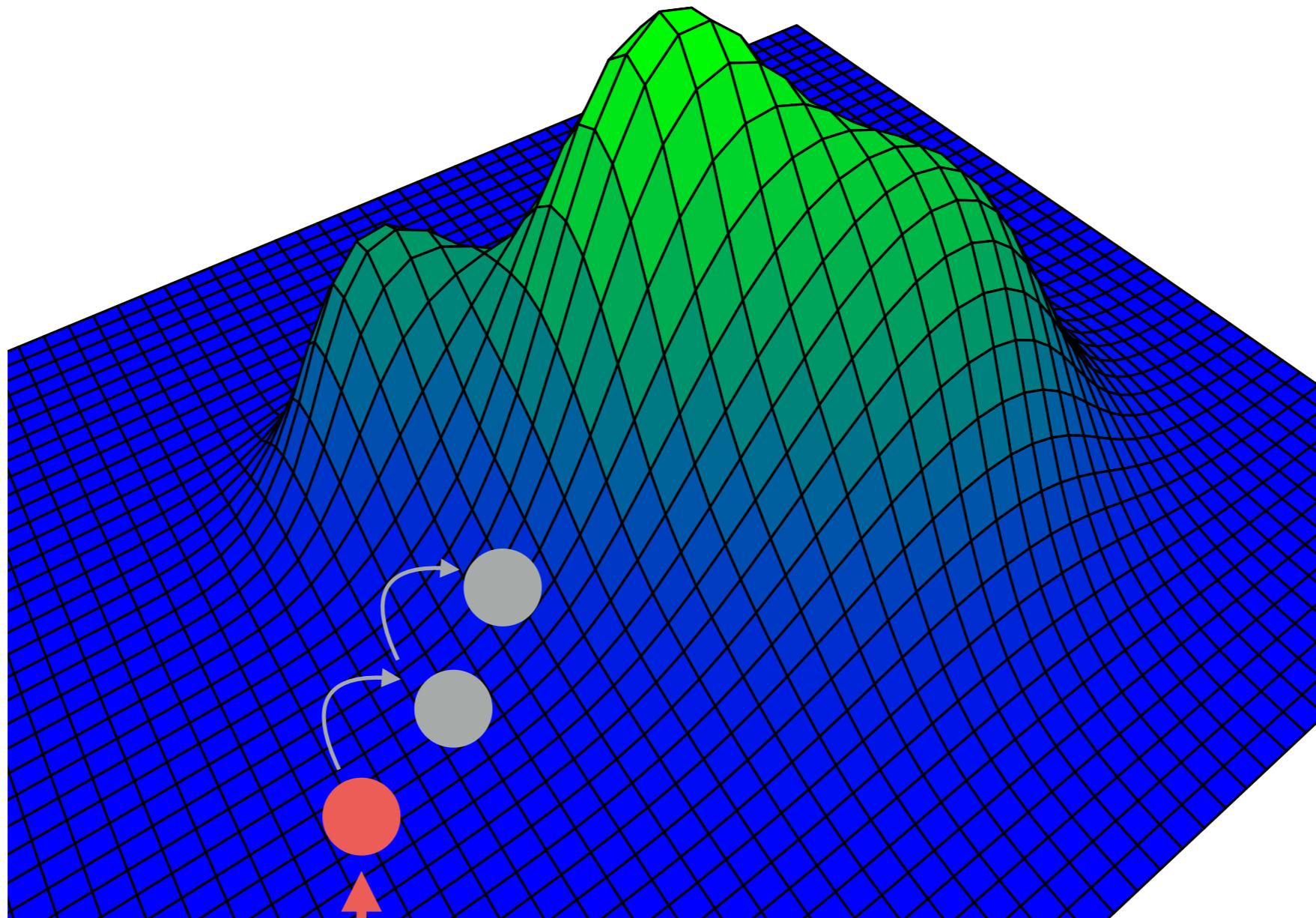
Starting tree

# Why is MCMC so slow? Traverse tree space



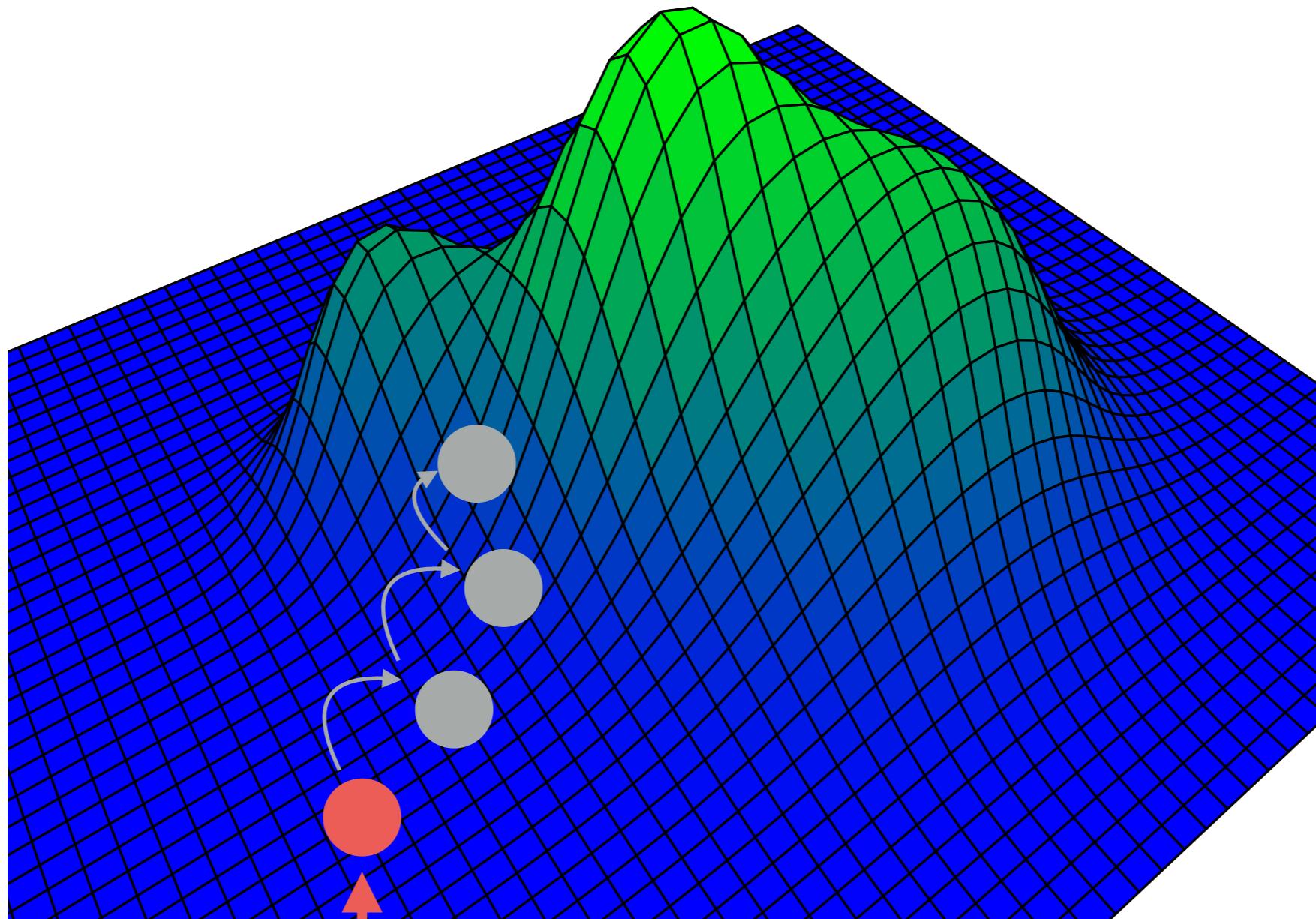
Starting tree

# Why is MCMC so slow? Traverse tree space



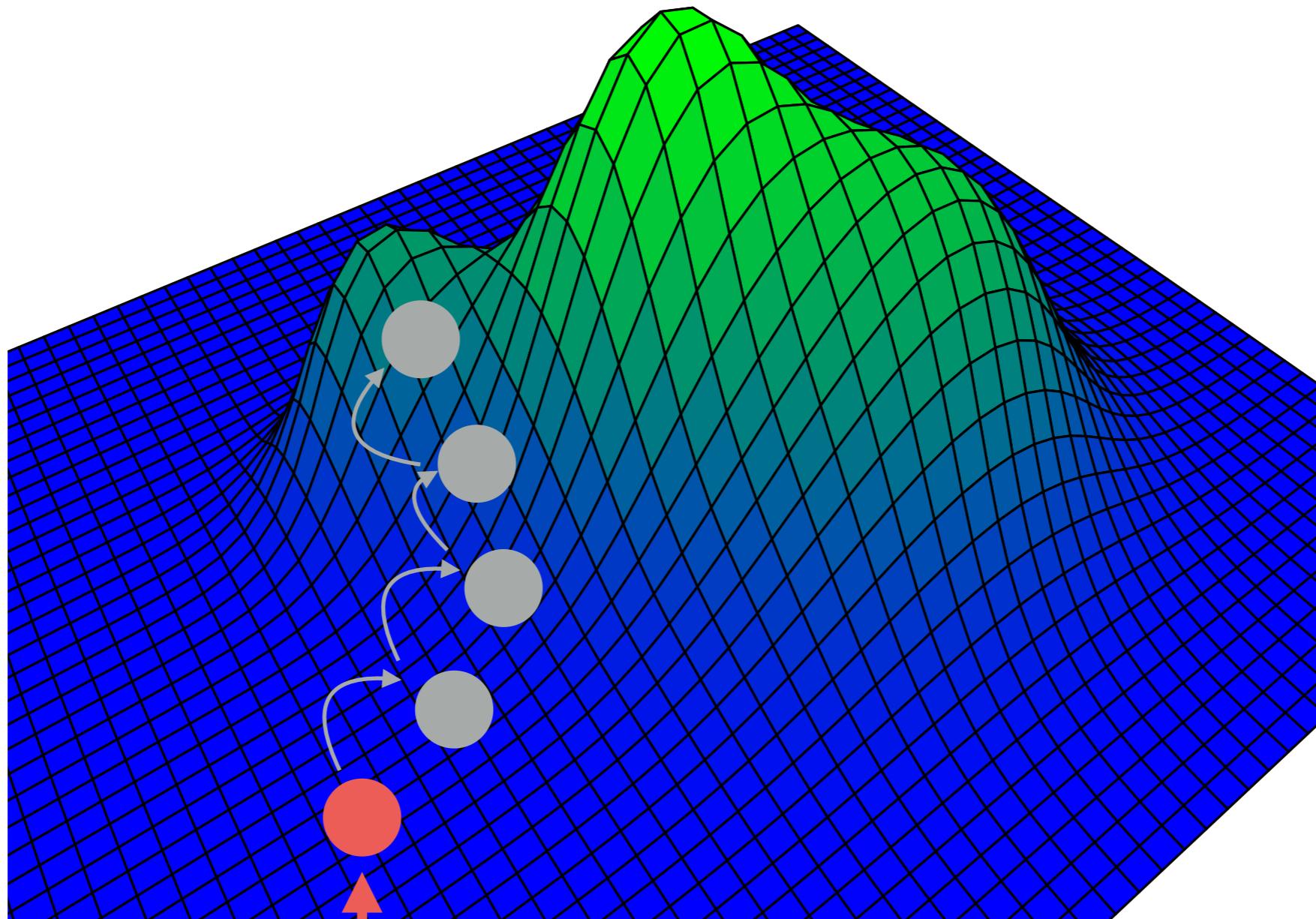
Starting tree

# Why is MCMC so slow? Traverse tree space



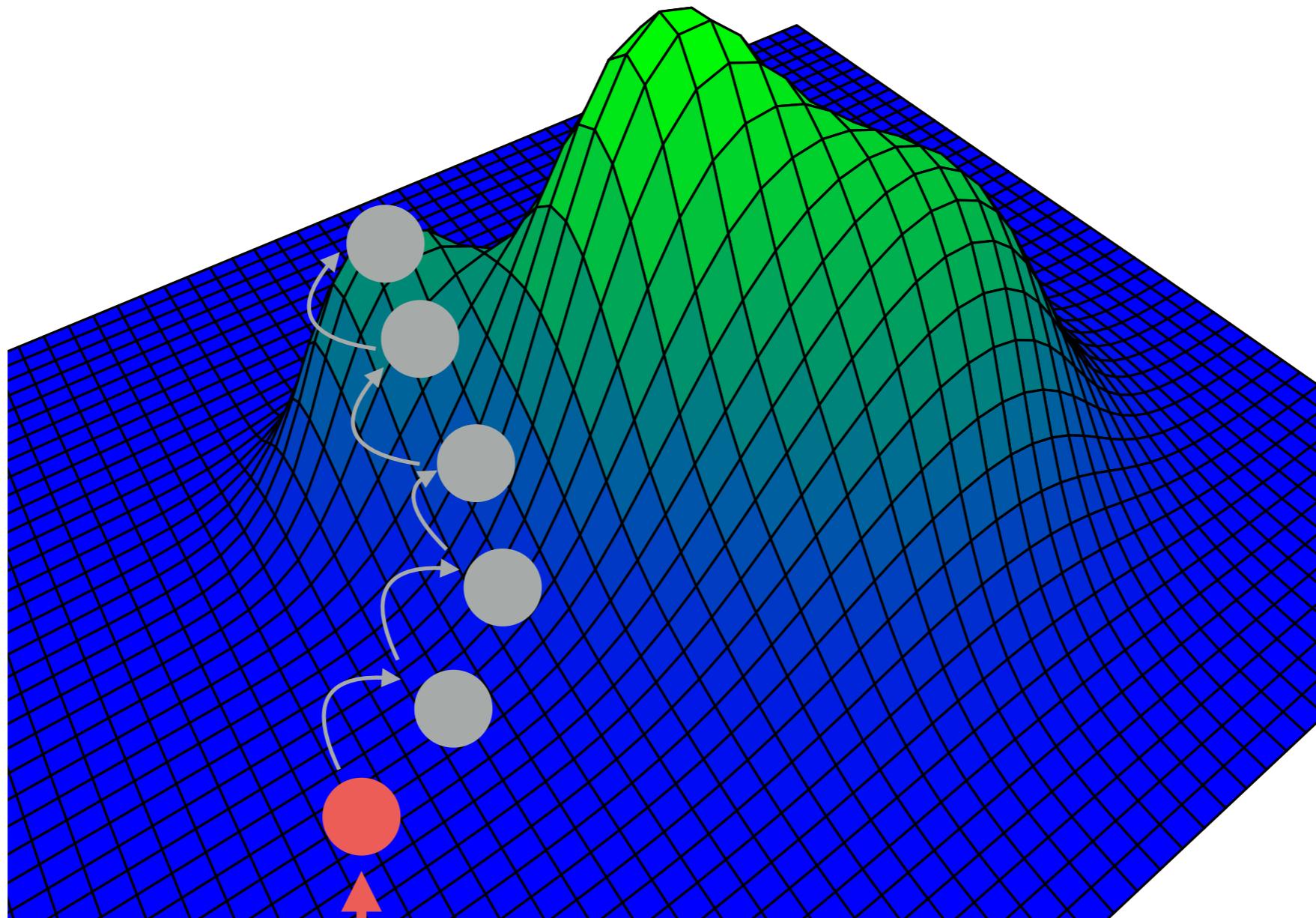
Starting tree

# Why is MCMC so slow? Traverse tree space



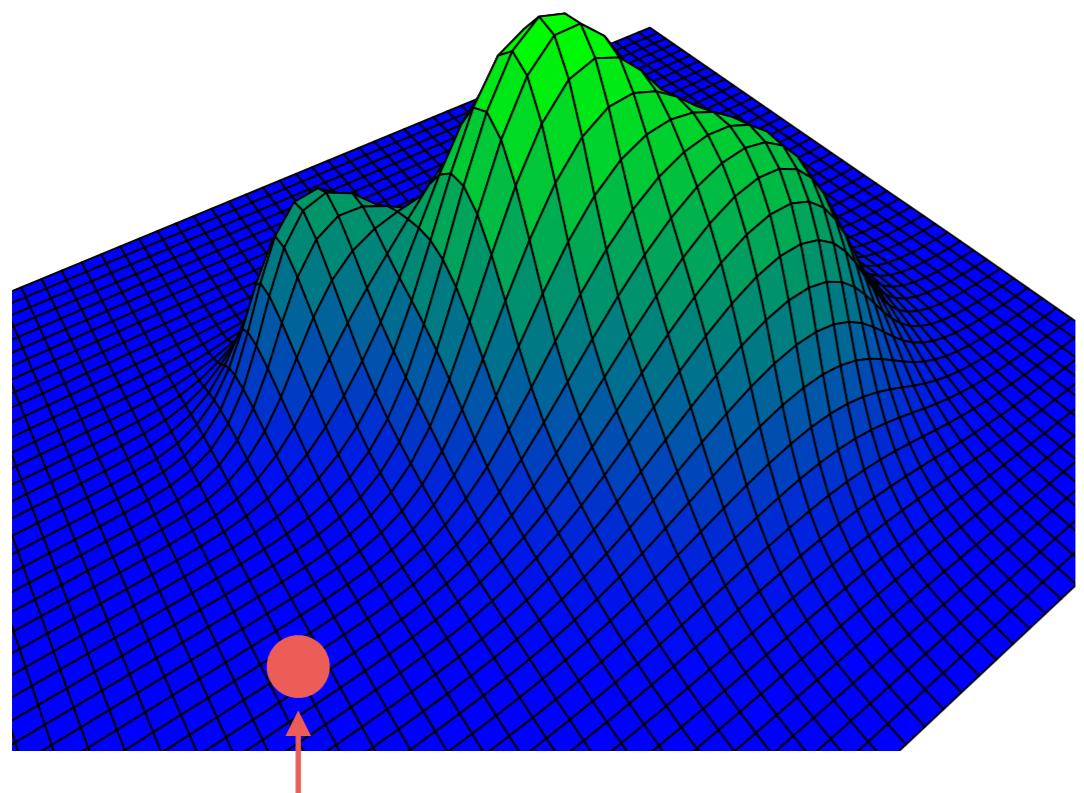
Starting tree

# Why is MCMC so slow? Traverse tree space

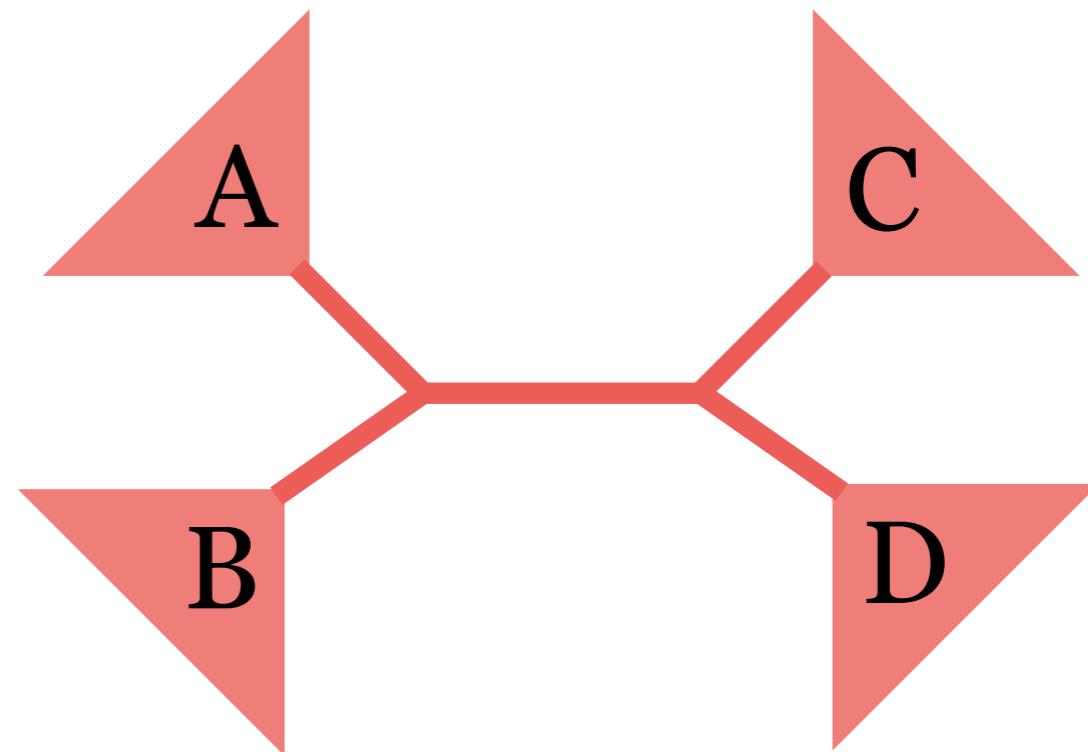


Starting tree

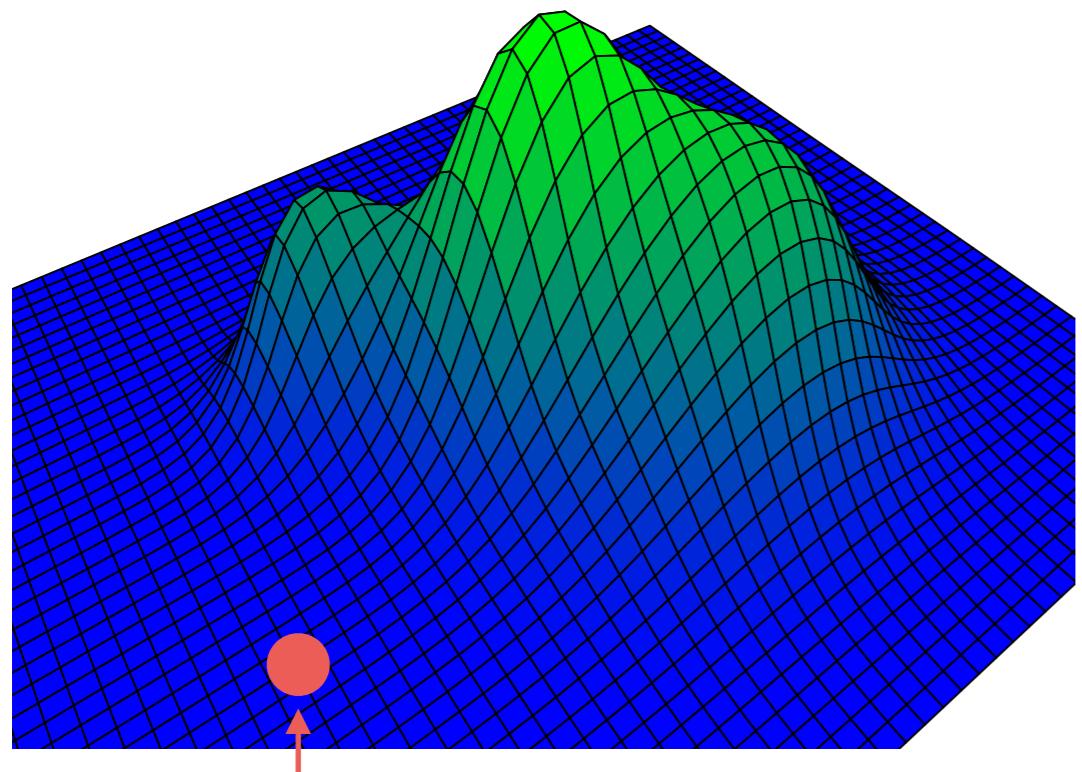
# Why is MCMC so slow? Traverse tree space



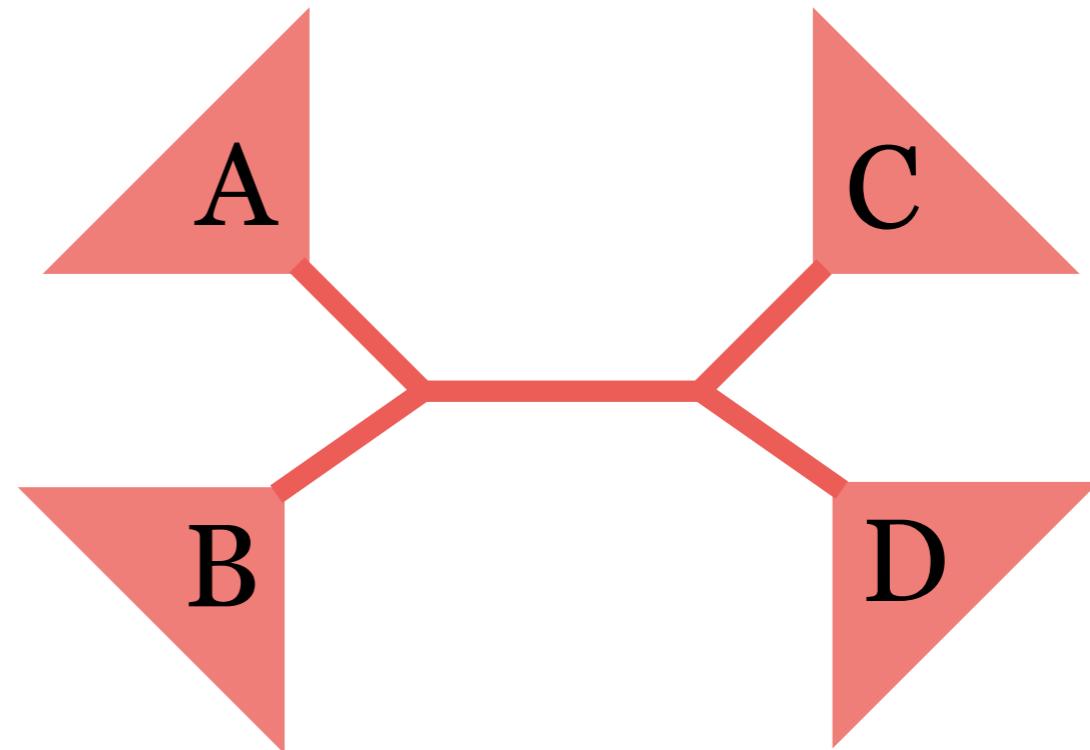
Starting tree



# Why is MCMC so slow? Traverse tree space

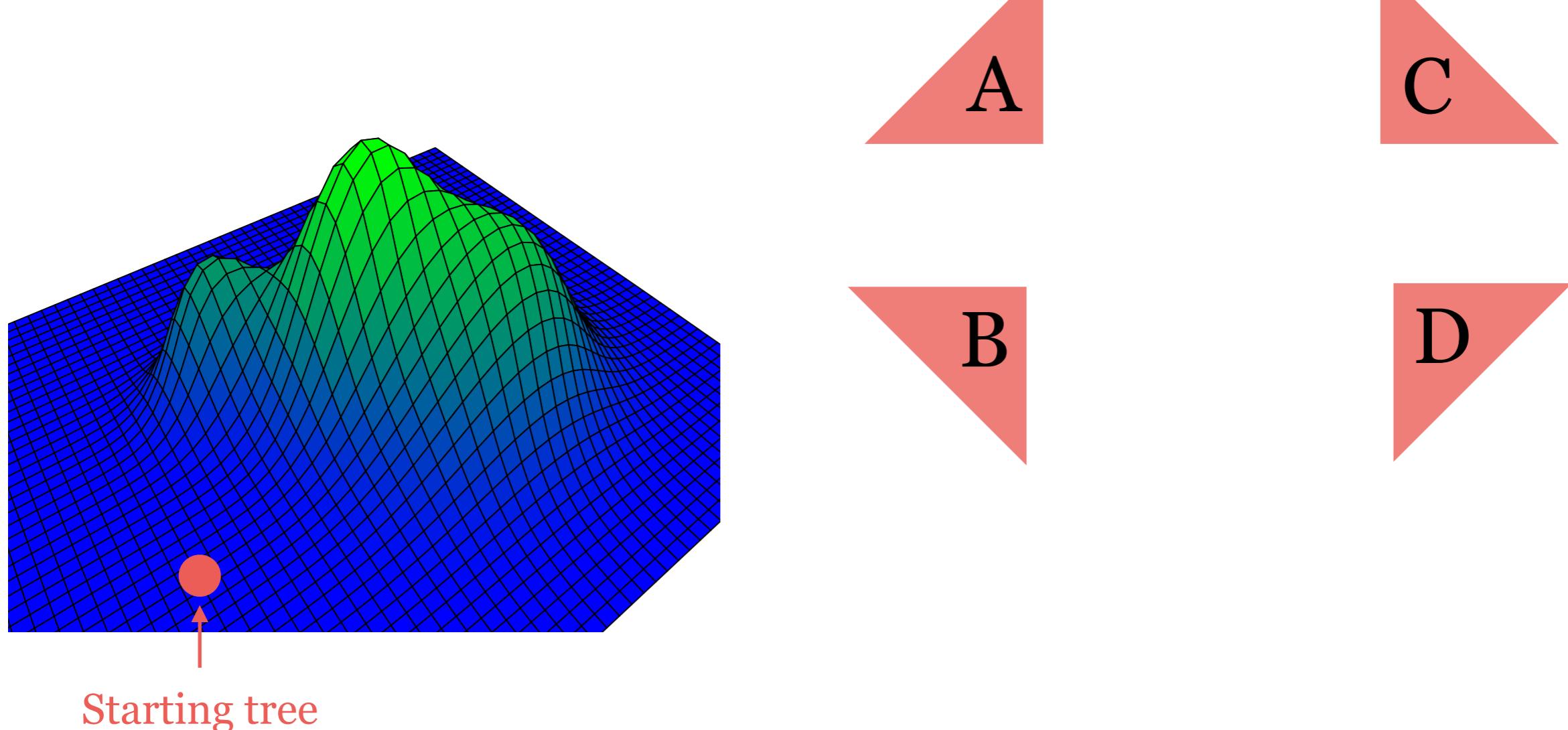


Starting tree



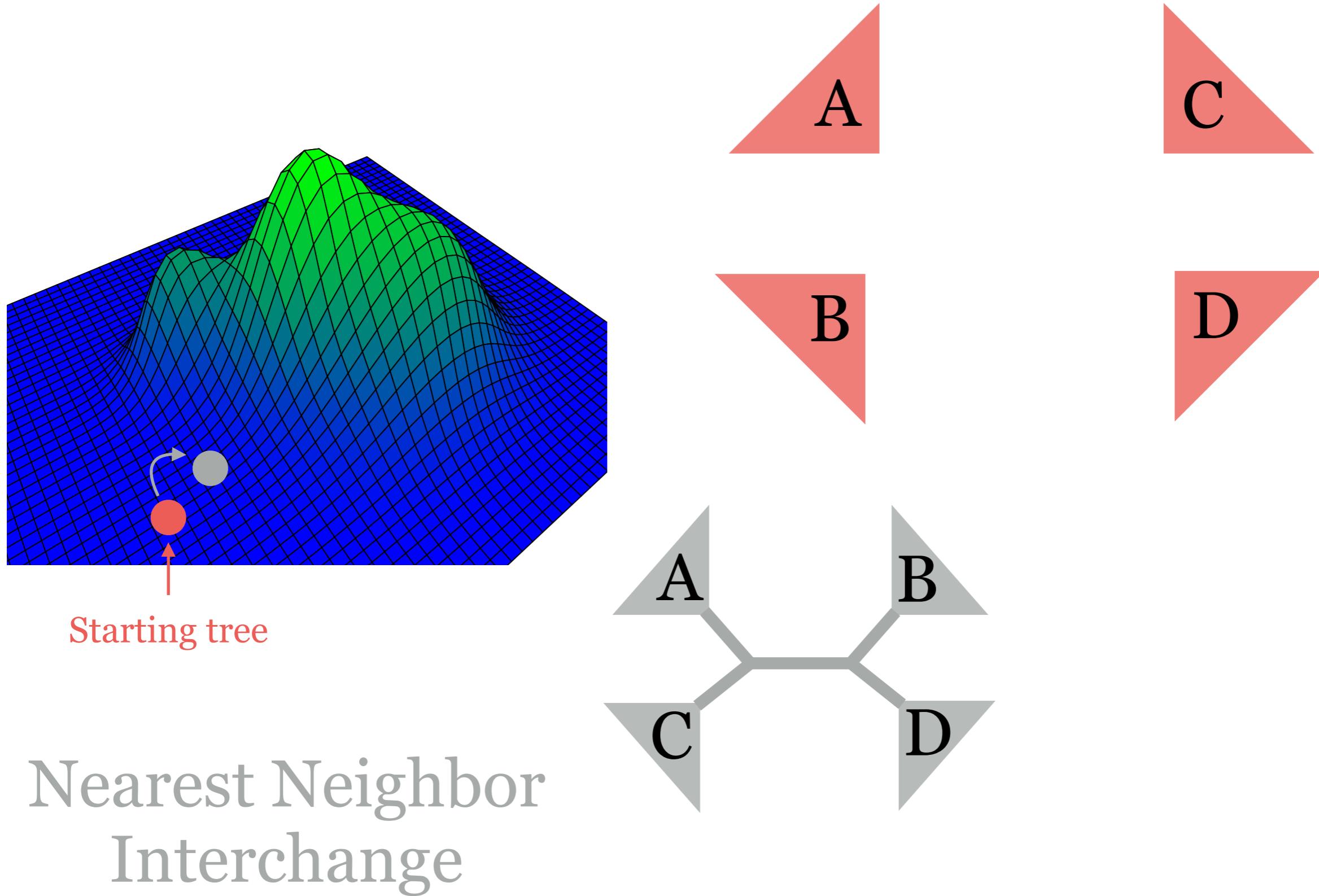
Nearest Neighbor  
Interchange

# Why is MCMC so slow? Traverse tree space

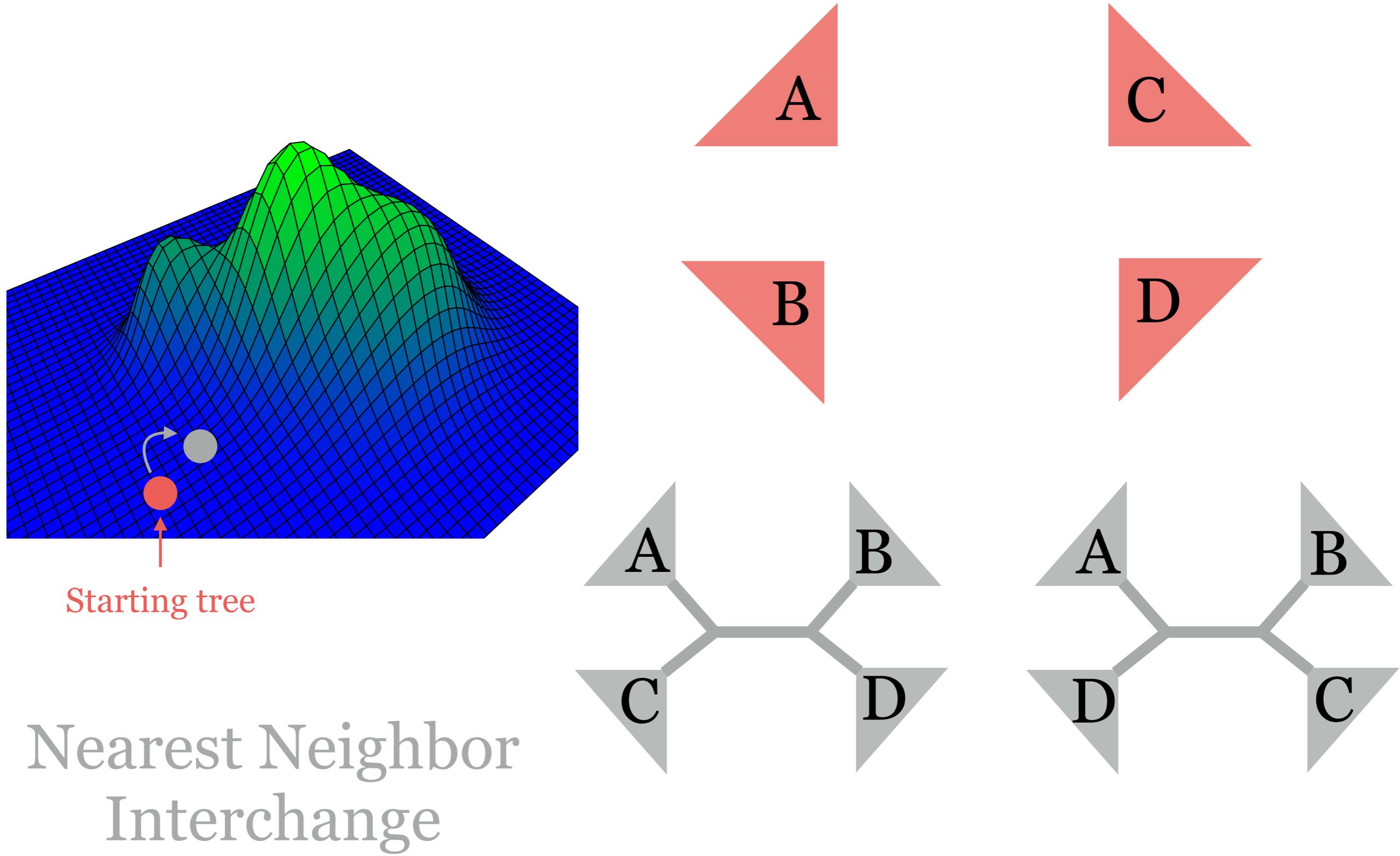


Nearest Neighbor  
Interchange

# Why is MCMC so slow? Traverse tree space



# Why is MCMC so slow? Traverse tree space



# Why is MCMC so slow?

Why is MCMC so slow? Tree space is huge

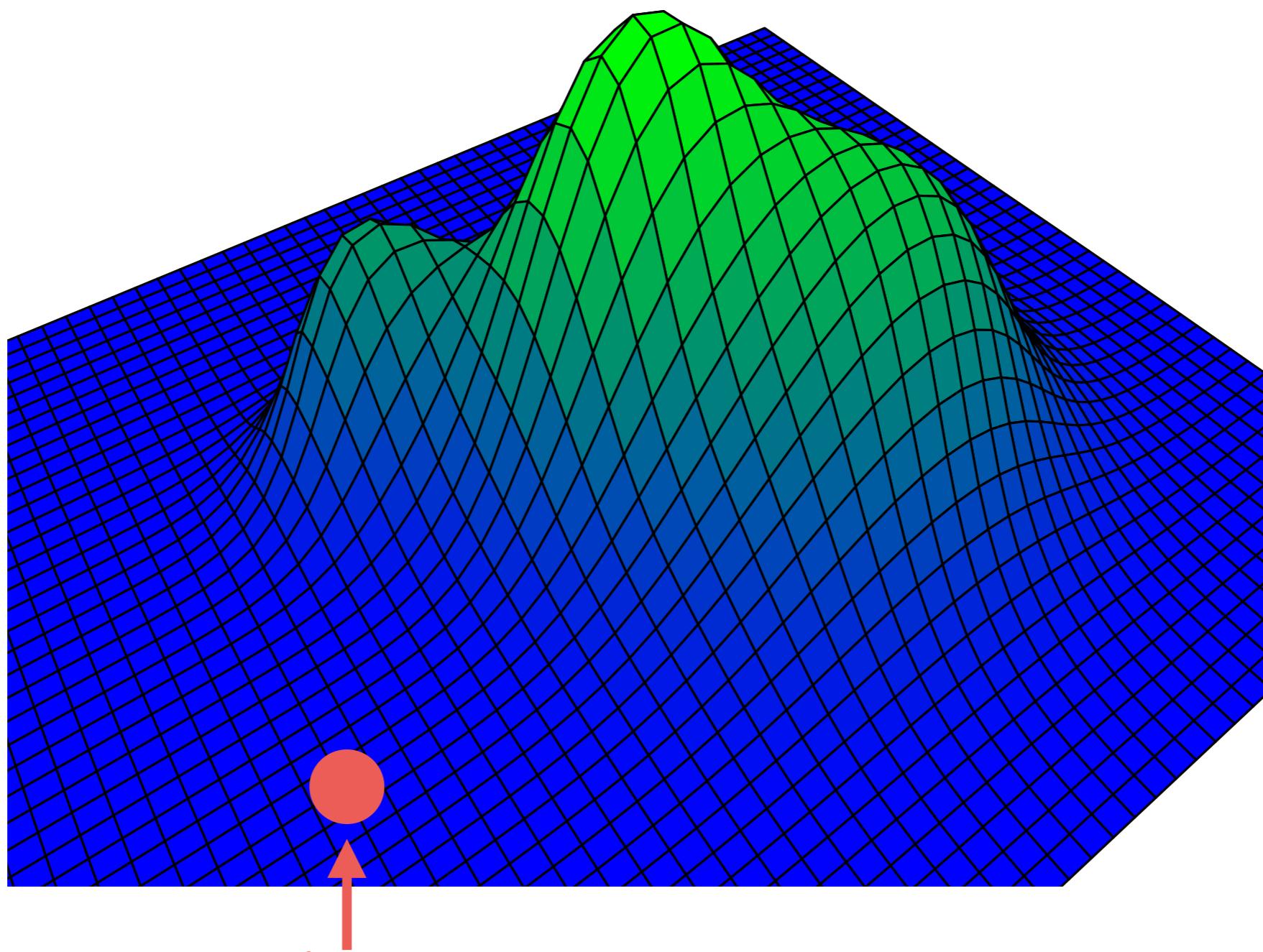
# Why is MCMC so slow? Tree space is huge

# Species	# Unrooted trees	# Rooted trees
1	1	1
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425
11	34,459,425	654,729,075
12	654,729,075	13,749,310,575
13	13,749,310,575	316,234,143,225
:	:	:
52	> # atoms in universe	

# Why is MCMC so slow?

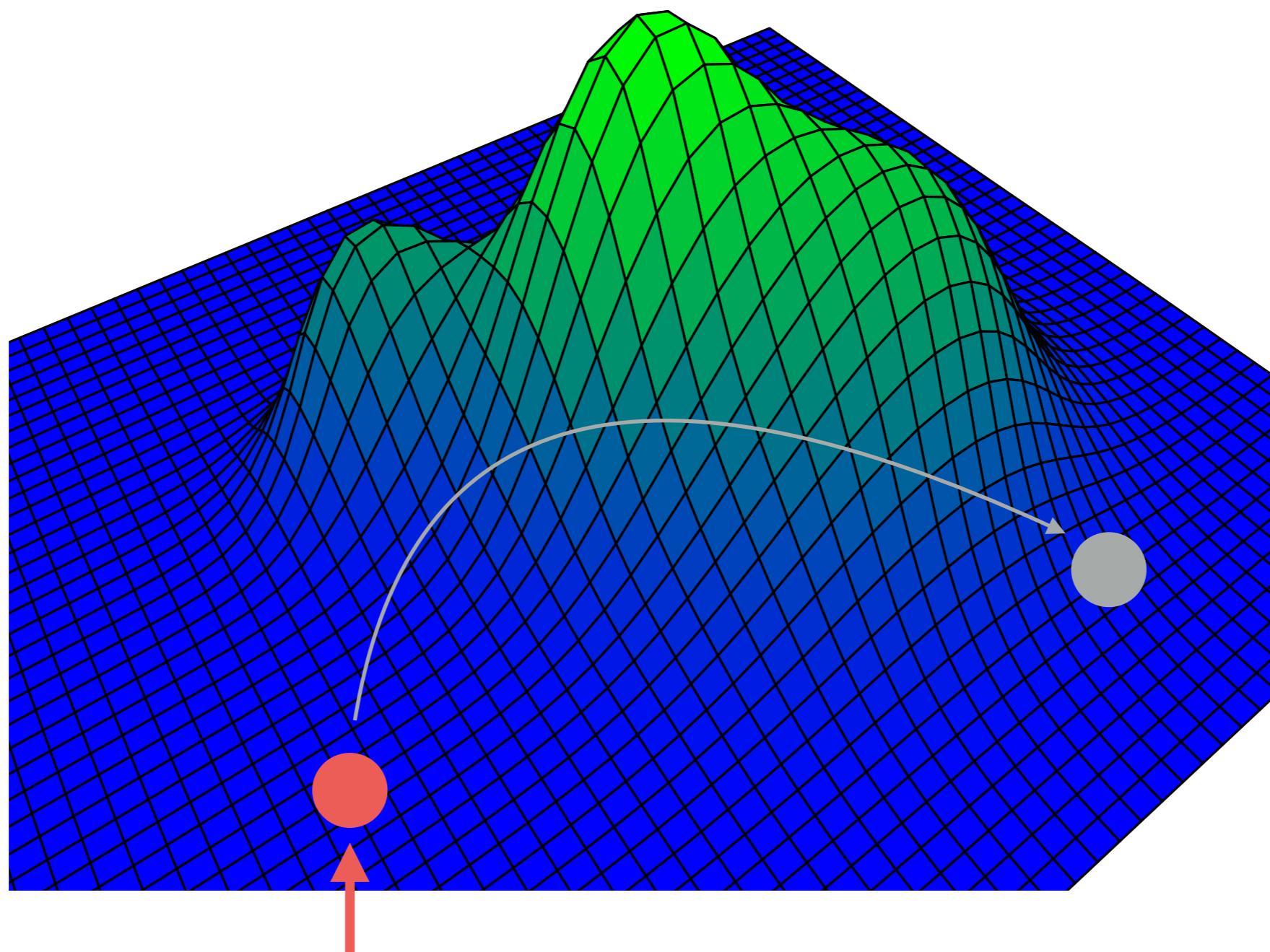
Why is MCMC so slow? Low acceptance of moves

# Why is MCMC so slow? Low acceptance of moves



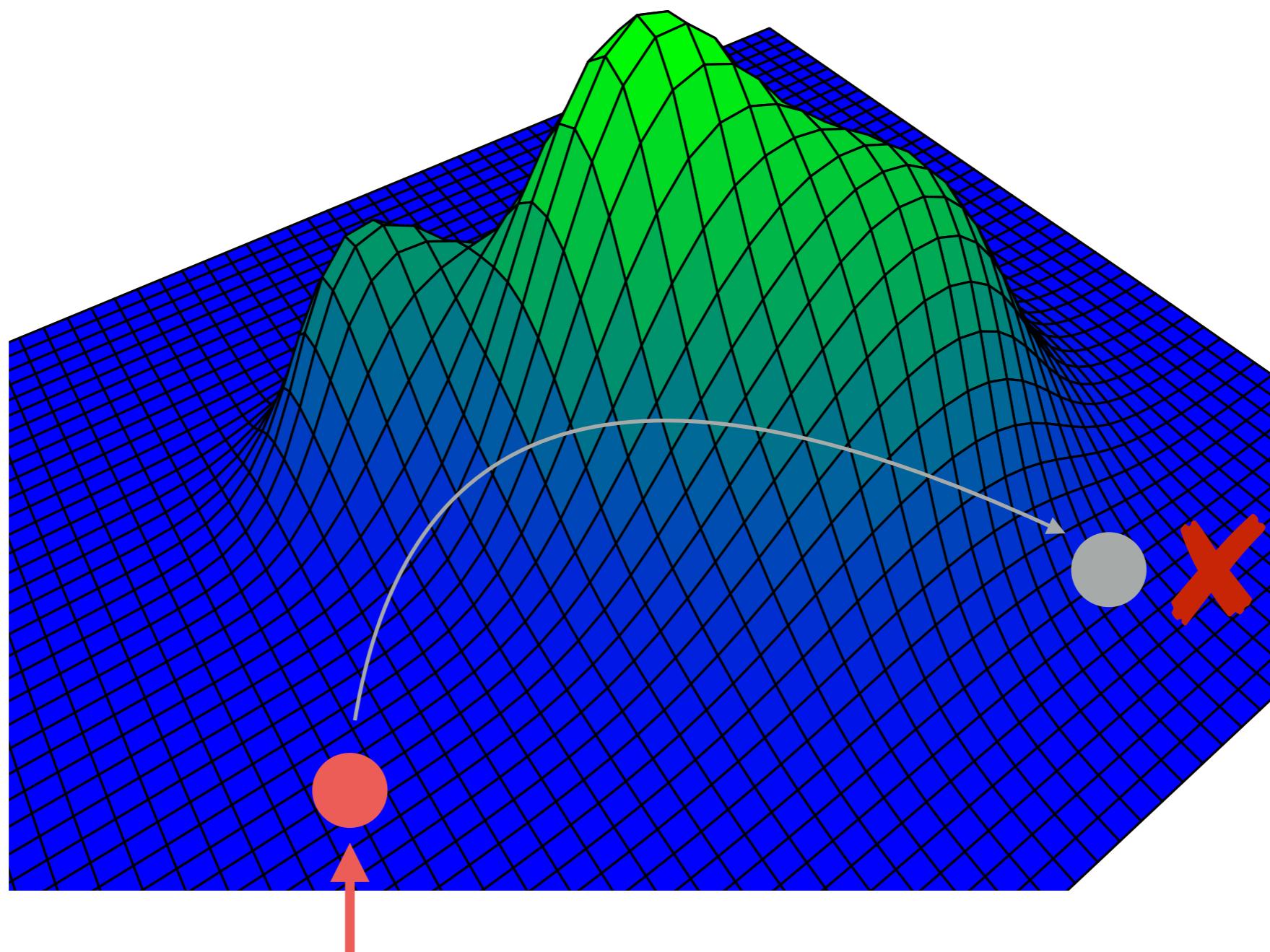
Starting tree

# Why is MCMC so slow? Low acceptance of moves



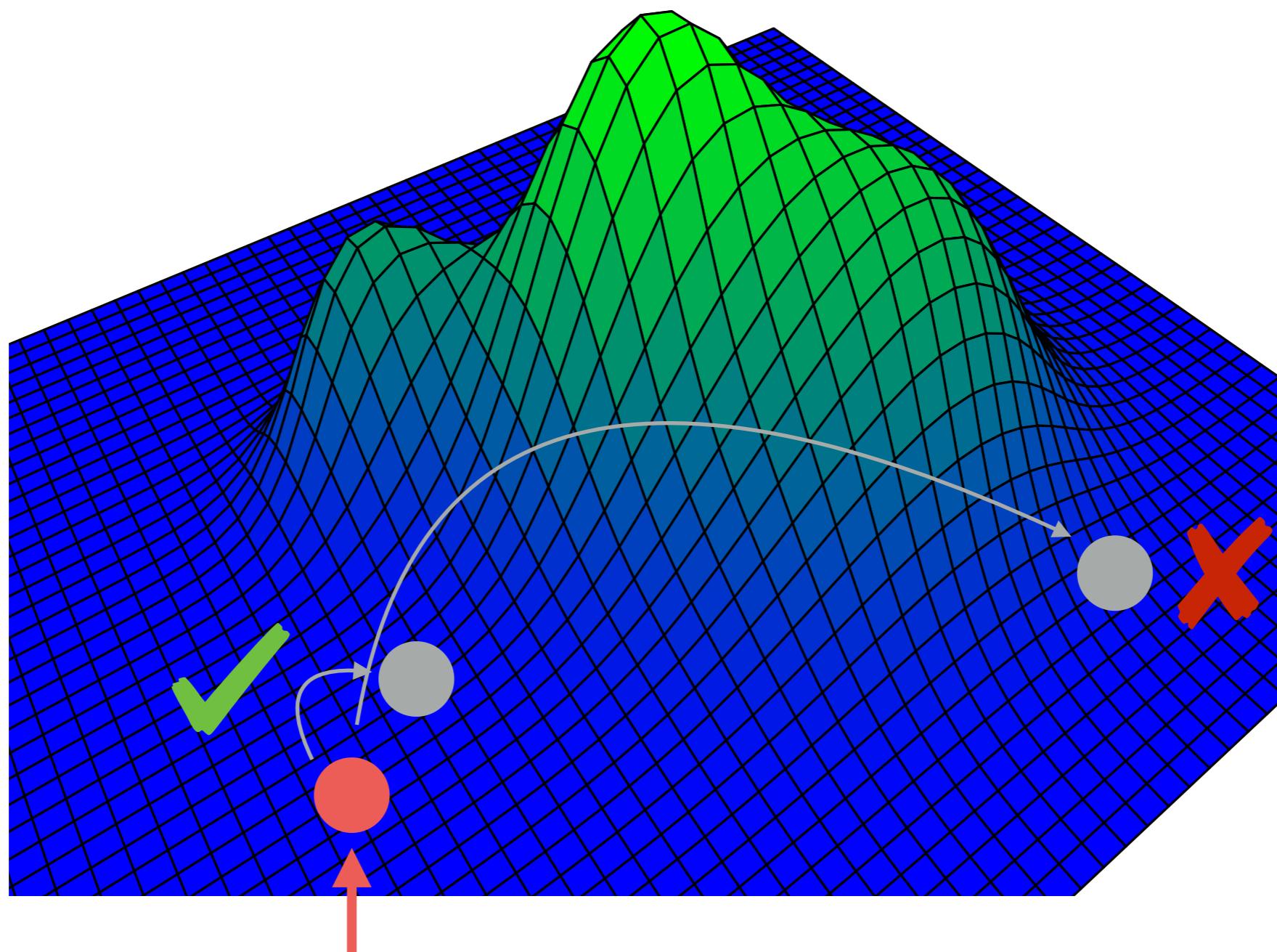
Starting tree

# Why is MCMC so slow? Low acceptance of moves



Starting tree

# Why is MCMC so slow? Low acceptance of moves



Starting tree

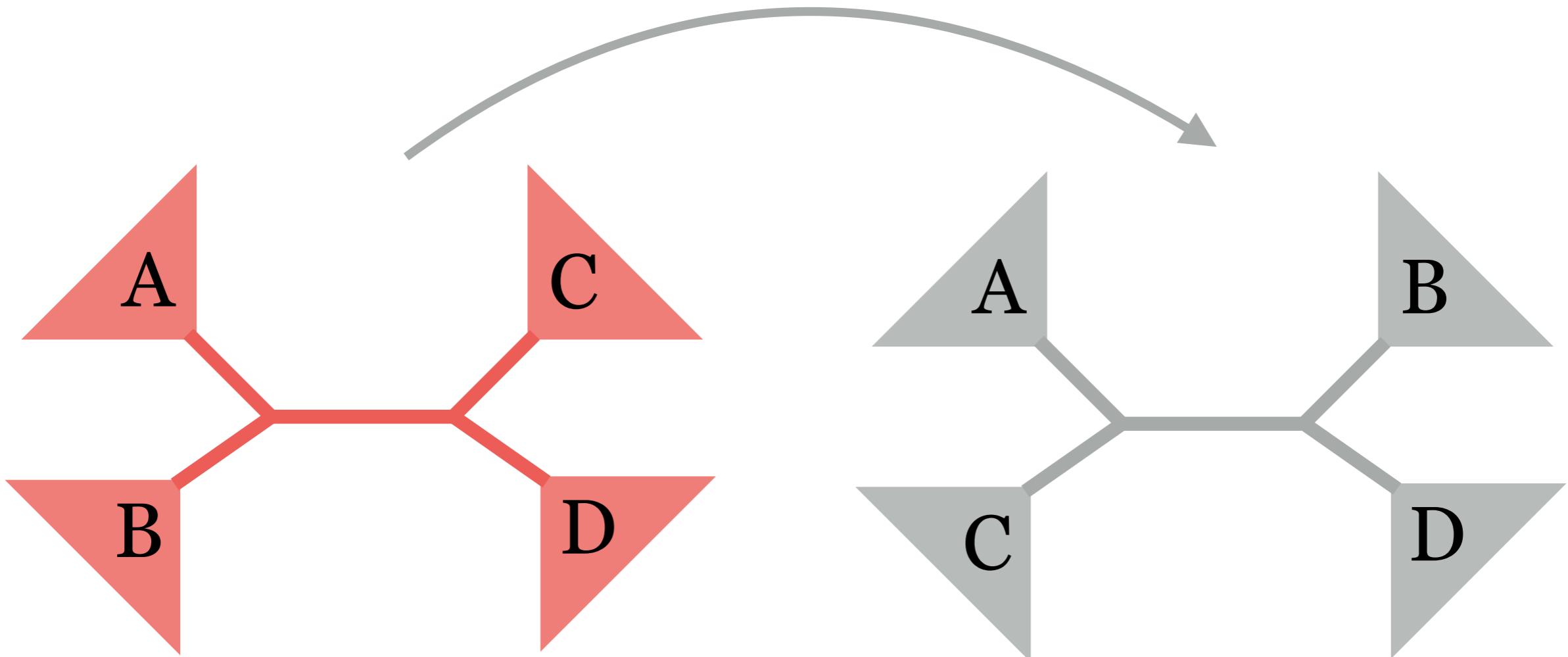
# Why is MCMC so slow?

# Why is MCMC so slow?

Small neighborhood  
implies very dependent  
sample

# Why is MCMC so slow?

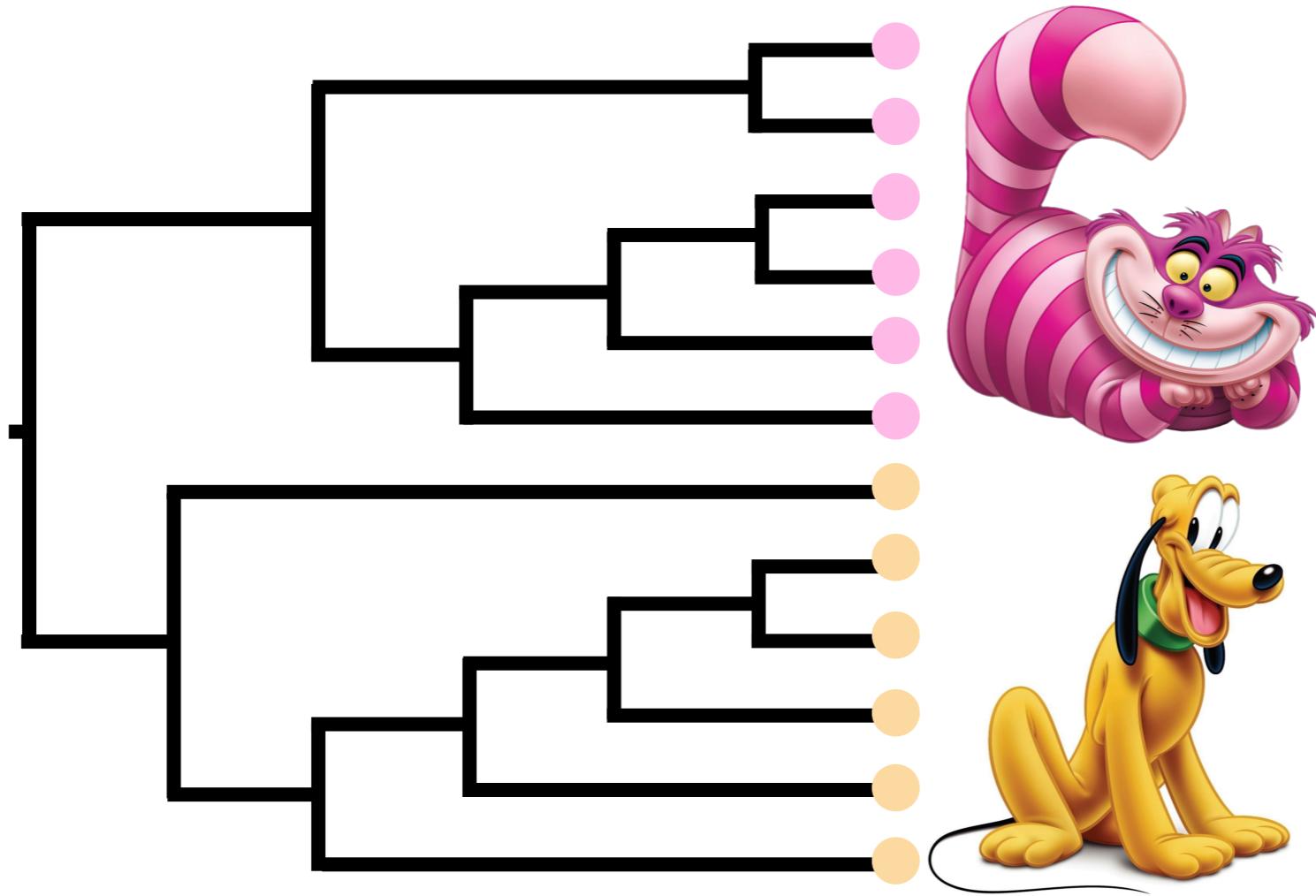
Small neighborhood  
implies very dependent  
sample



# Why is MCMC so slow?

- 1) Huge tree space size
- 2) Low acceptance of moves unless small neighborhood
- 3) Small neighborhood implies very dependent sample, which means small effective sample size

We need a gigantic chain because the space is huge and we are making tiny moves



12 taxa *Carnivora*

MCMC efficiency  $\sim 0.025\%$

(250 from 1 million post-burnin generations)

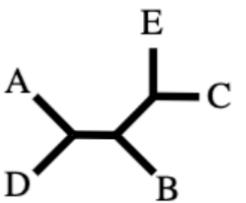
ESS  
↑

# Priors

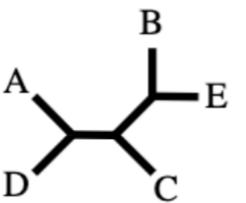
# Common Priors

- **Discrete uniform** for topologies
  - exceptions becoming more common
- **Beta** for proportions
- **Gamma** or **Log-normal** for branch lengths and other parameters with support  $[0, \infty)$ 
  - Exponential is common special case of the gamma distribution
- **Dirichlet** for state frequencies and GTR relative rates

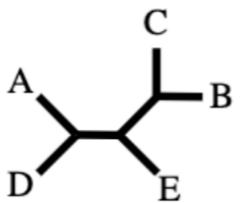
# Discrete Uniform distribution for topologies



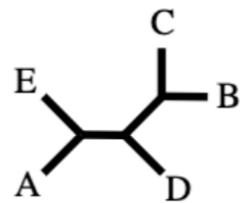
15



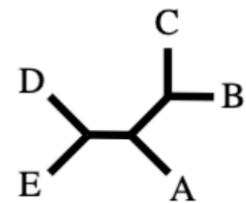
1  
15



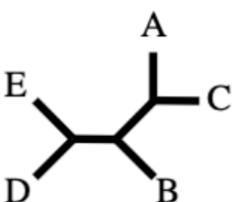
1  
15



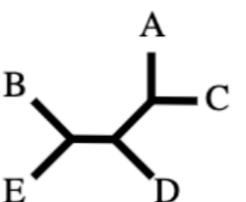
15



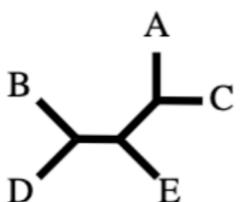
1  
15



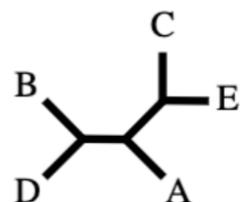
15



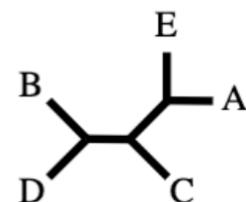
1  
15



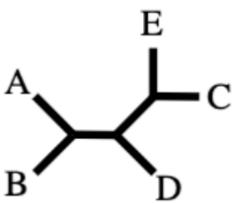
15



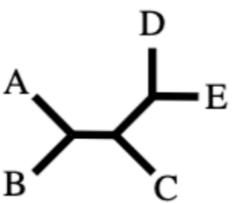
1  
15



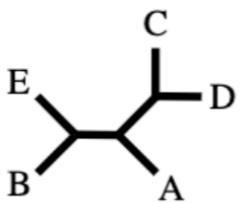
1  
15



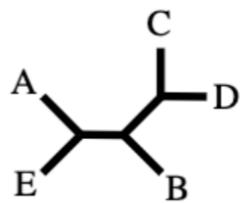
1  
15



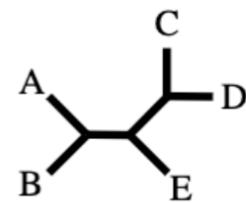
1  
15



15

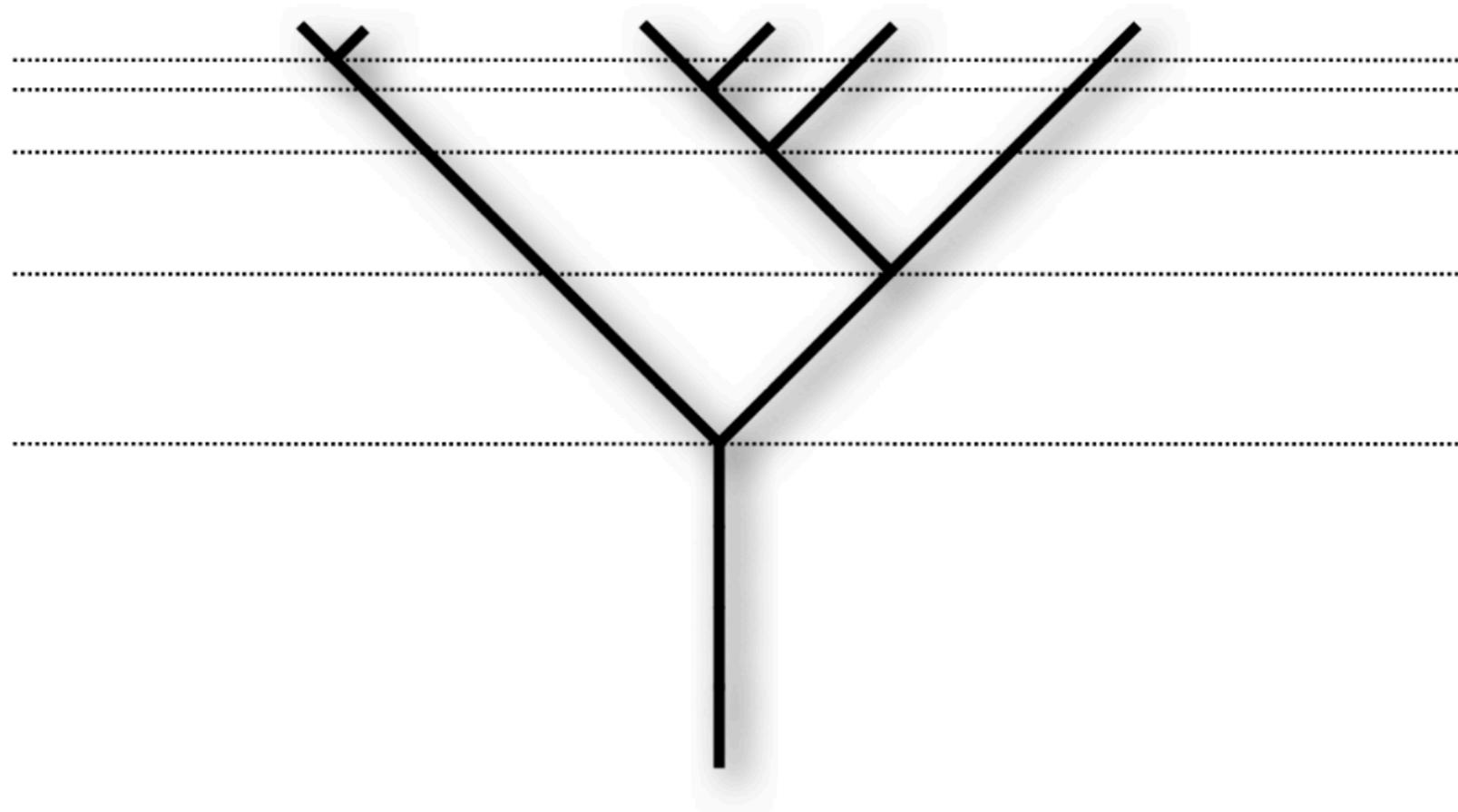


1  
15



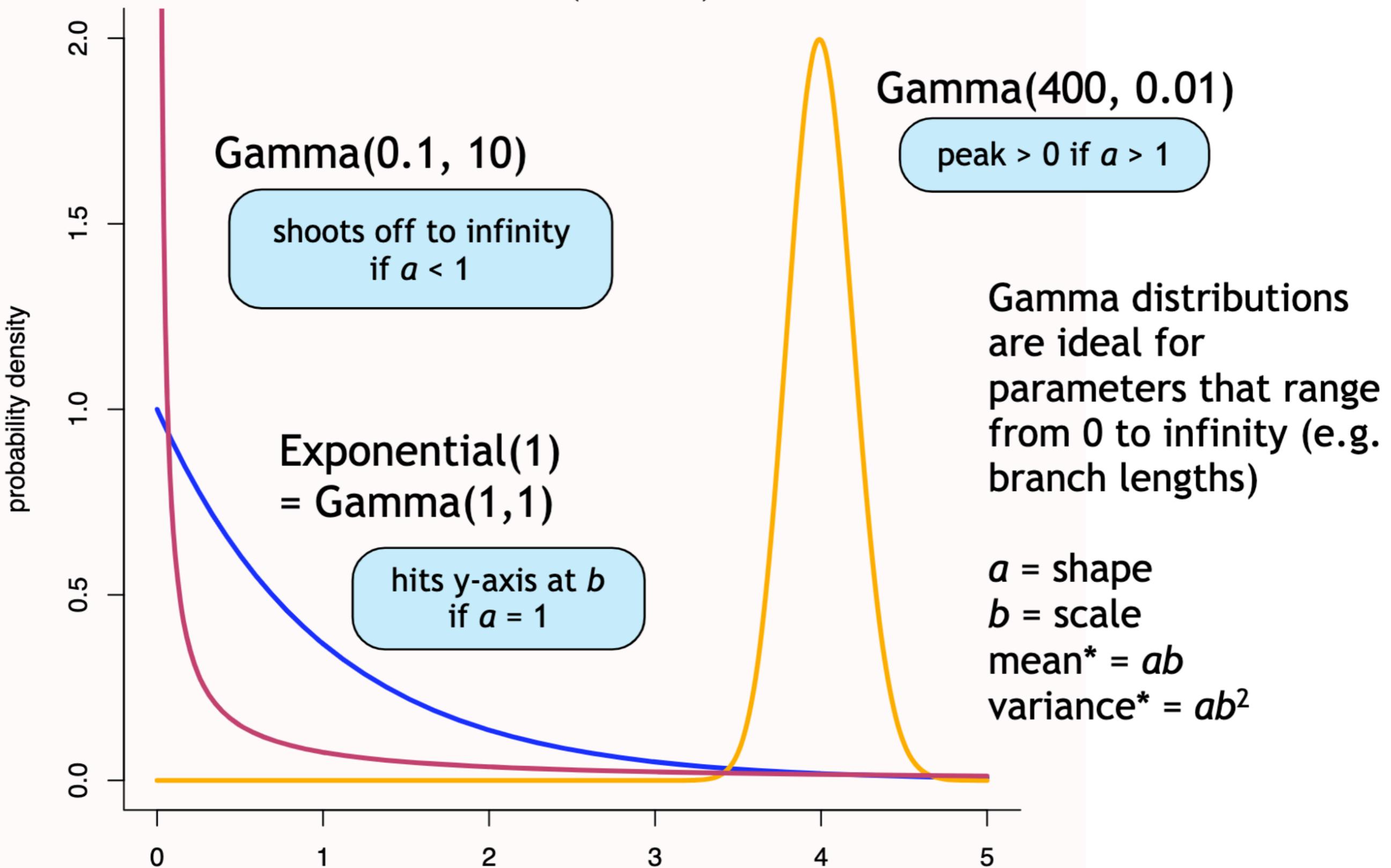
1  
15

# Yule model provides joint prior for both topology and divergence times



The rate of speciation under the Yule model ( $\lambda$ ) is constant and applies equally and independently to each lineage. Thus, speciation events get closer together in time as the tree grows because more lineages are available to speciate.

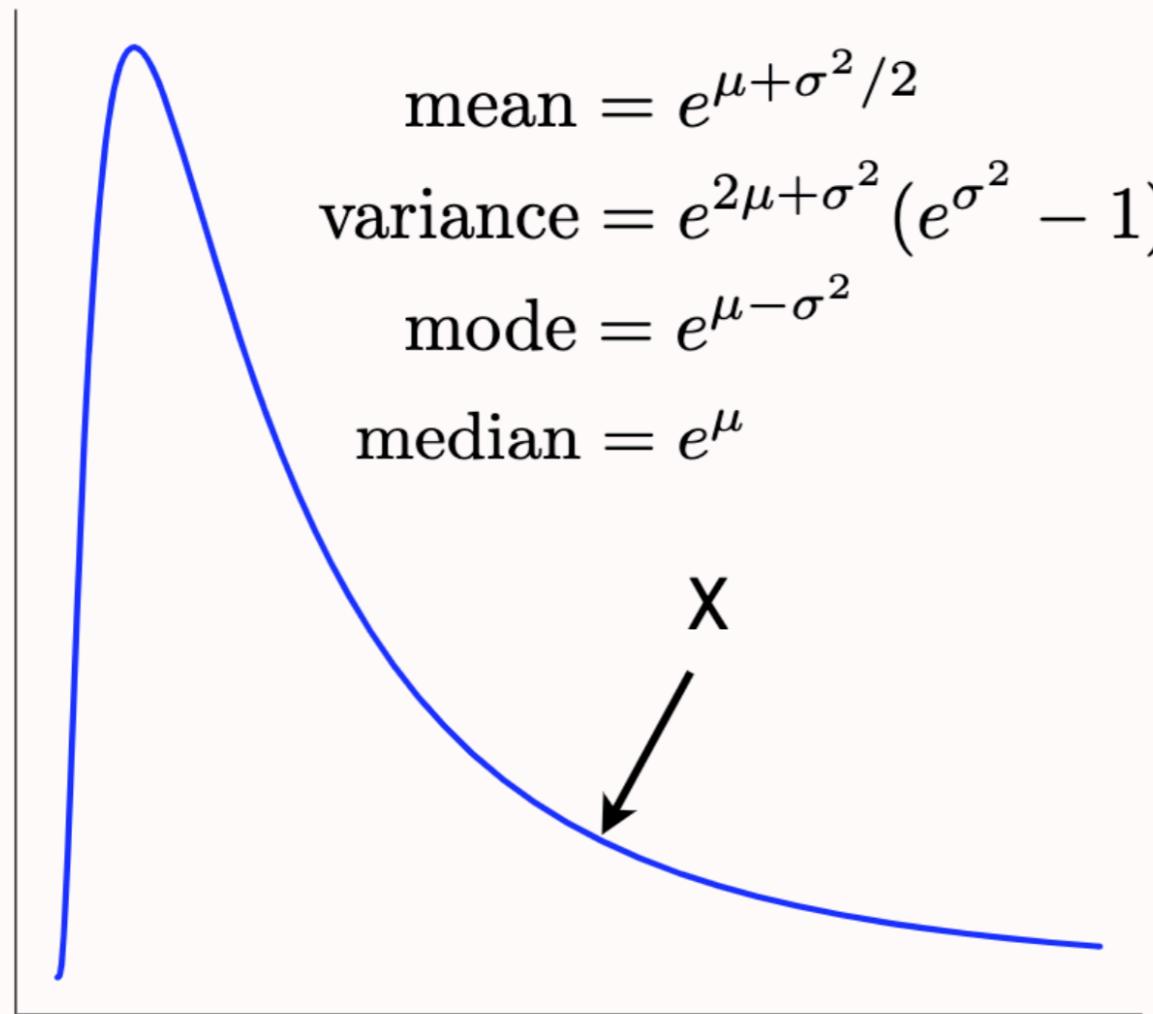
# Gamma( $a, b$ ) distributions



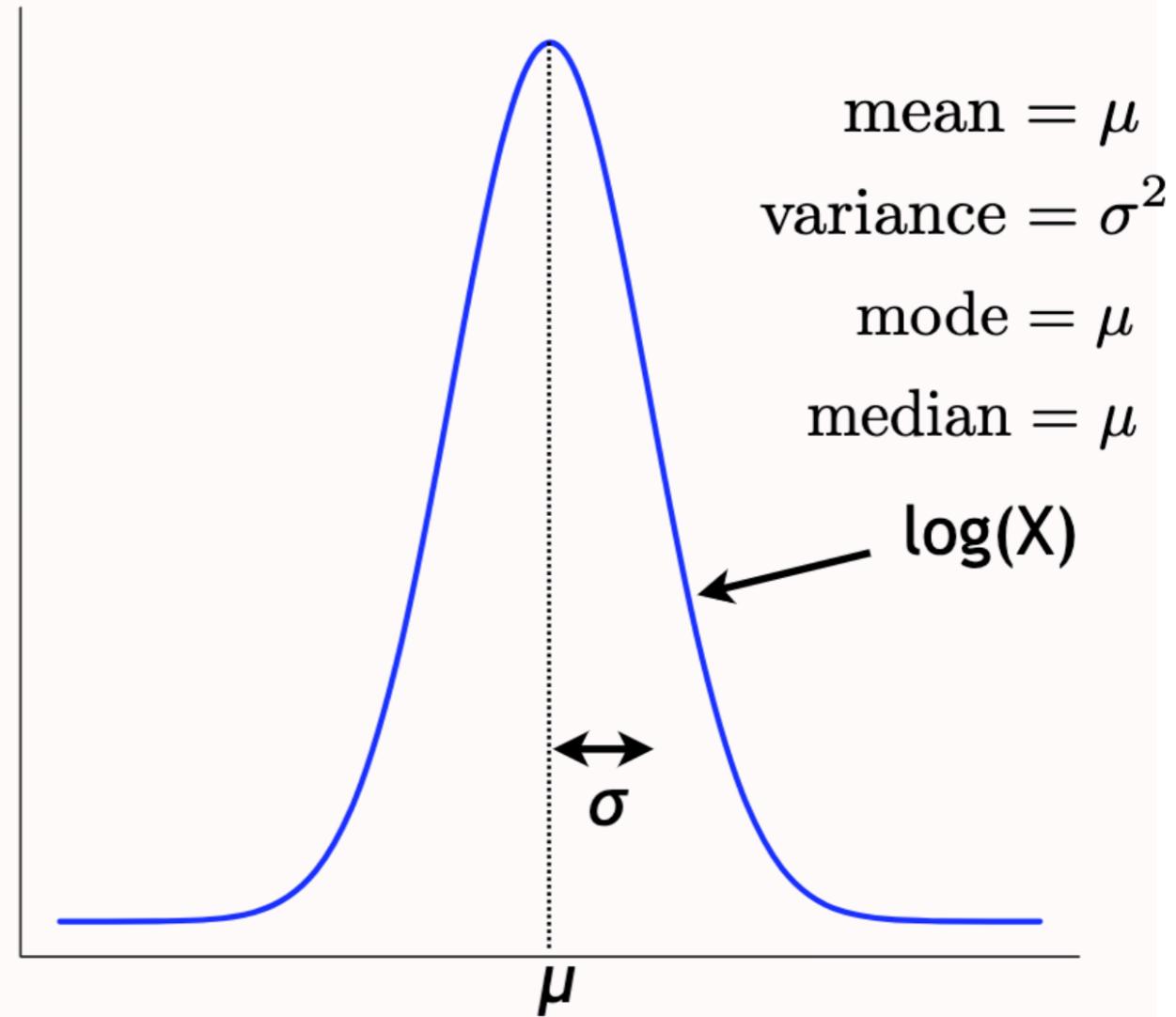
\*Note: be aware that in many papers the Gamma distribution is defined such that the second (scale) parameter is the *inverse* of the value  $b$  used in this slide! In this case, the mean and variance would be  $a/b$  and  $a/b^2$ , respectively.

# Log-normal distribution

If  $X$  is log-normal with *parameters*  $\mu$  and  $\sigma$ ...



...then  $\log(X)$  is normal with *mean*  $\mu$  and *standard deviation*  $\sigma$ .

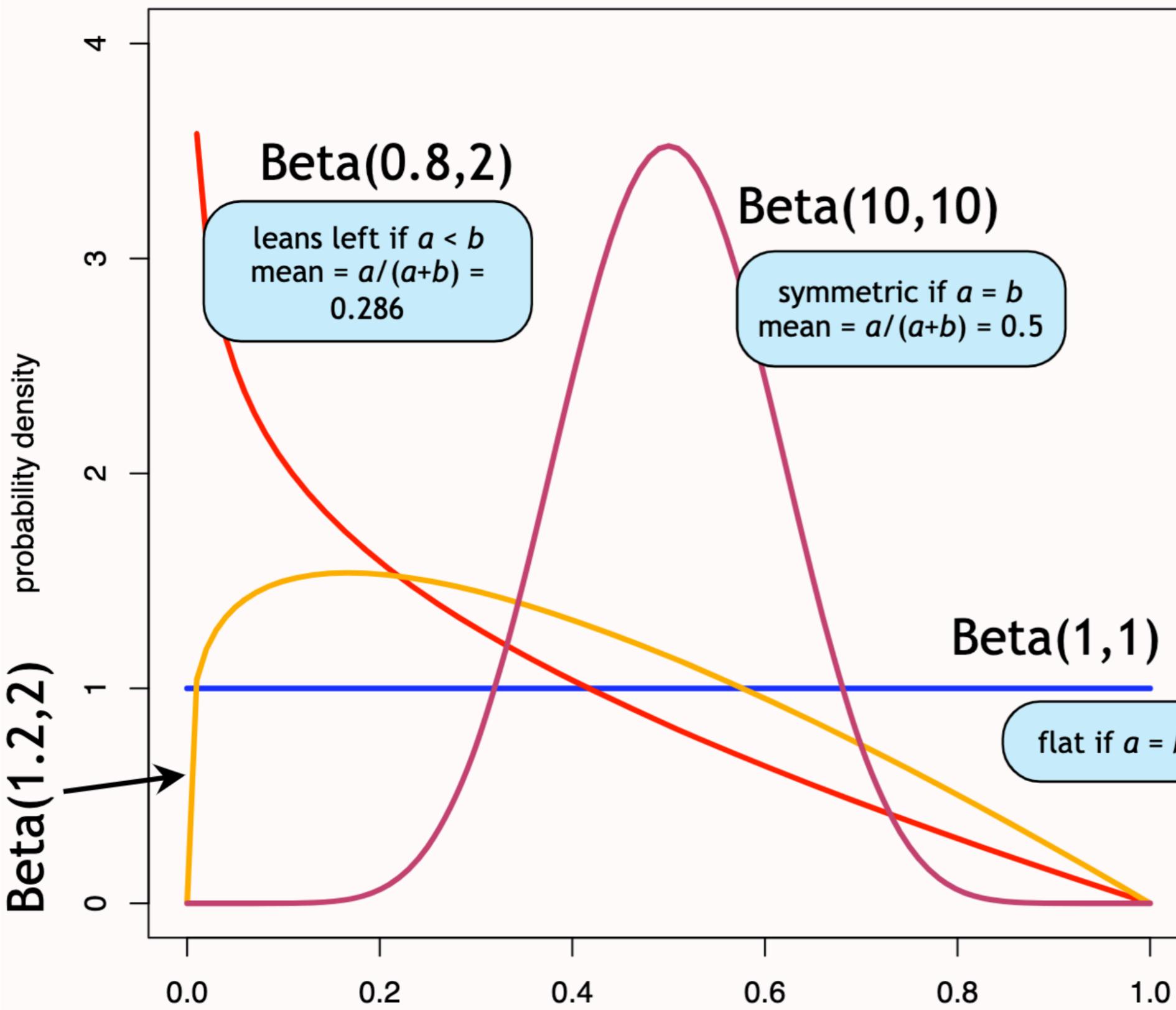


**Important:**  $\mu$  and  $\sigma$  do **not** represent the mean and standard deviation of  $X$ : they are the mean and standard deviation of  $\log(X)$ !

To choose  $\mu$  and  $\sigma$  to yield a particular mean ( $m$ ) and variance ( $v$ ) for  $X$ , use these formulas:

$$\mu = \log(m^2) - \log(m) - \frac{\log(v + m^2) - \log(m^2)}{2}$$
$$\sigma^2 = \log(v + m^2) - \log(m^2)$$

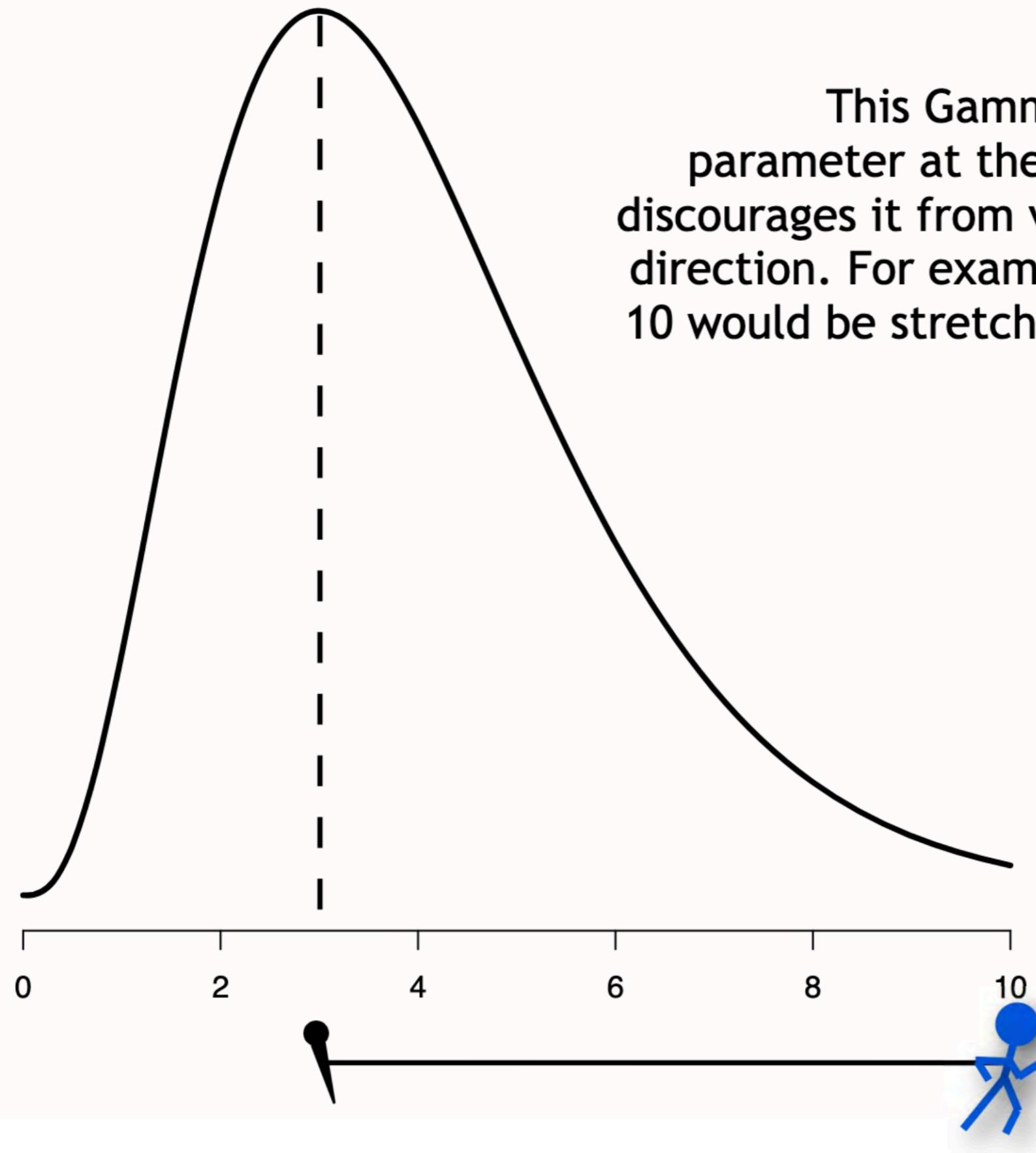
# Beta( $a,b$ ) gallery



Beta distributions are appropriate for proportions, which are constrained to the interval [0,1].

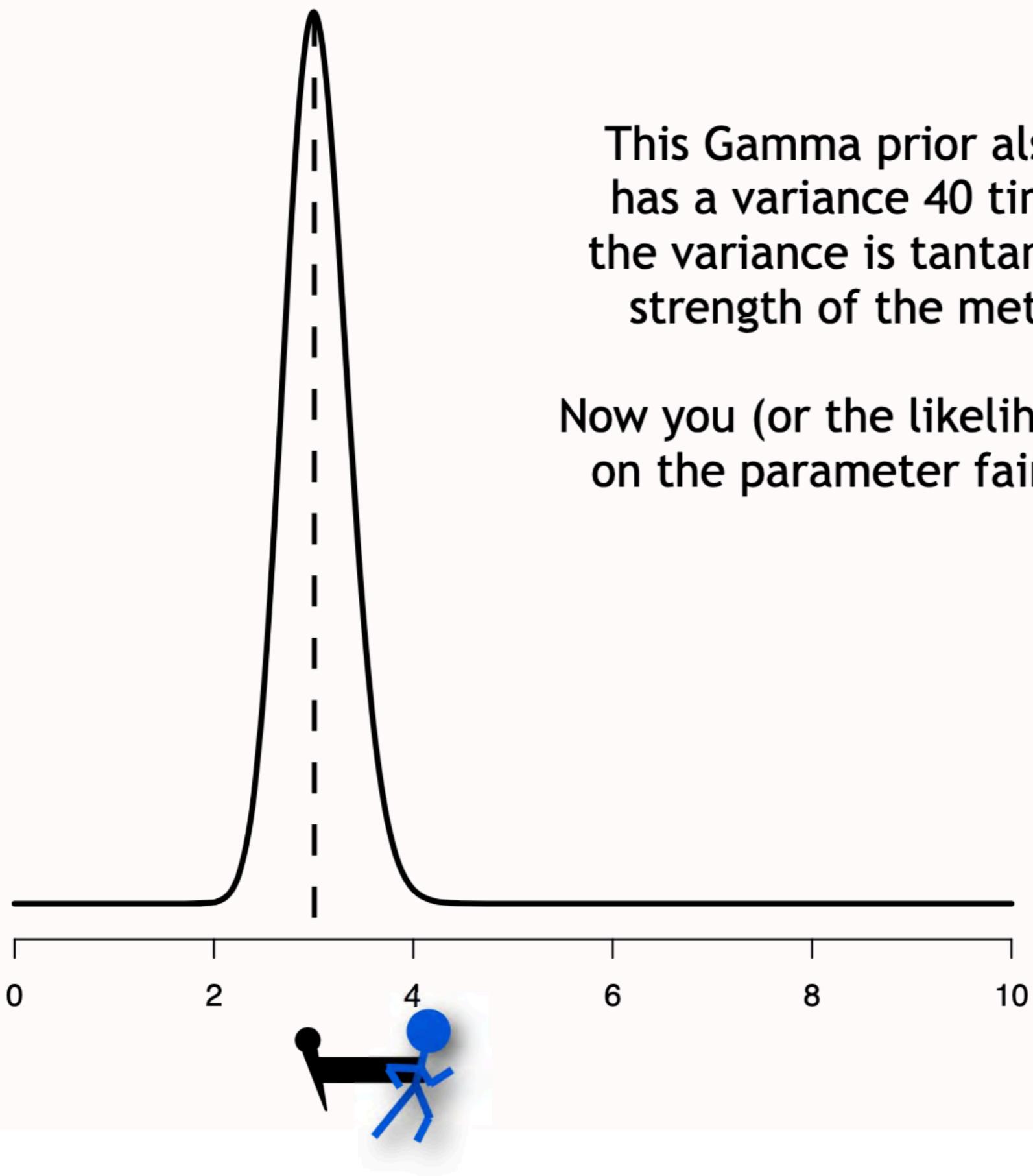
$$\text{mean} = a/(a+b)$$
$$\text{variance} = ab/[(a+b)^2(a+b+1)]$$

**Non-informative prior**



This  $\text{Gamma}(4,1)$  prior ties down its parameter at the mode, which is at 3, and discourages it from venturing too far in either direction. For example, a parameter value of 10 would be stretching the rubber band fairly tightly

The mode of a  $\text{Gamma}(a,b)$  distribution is  $(a-1)b$  (assuming  $a > 1$ )

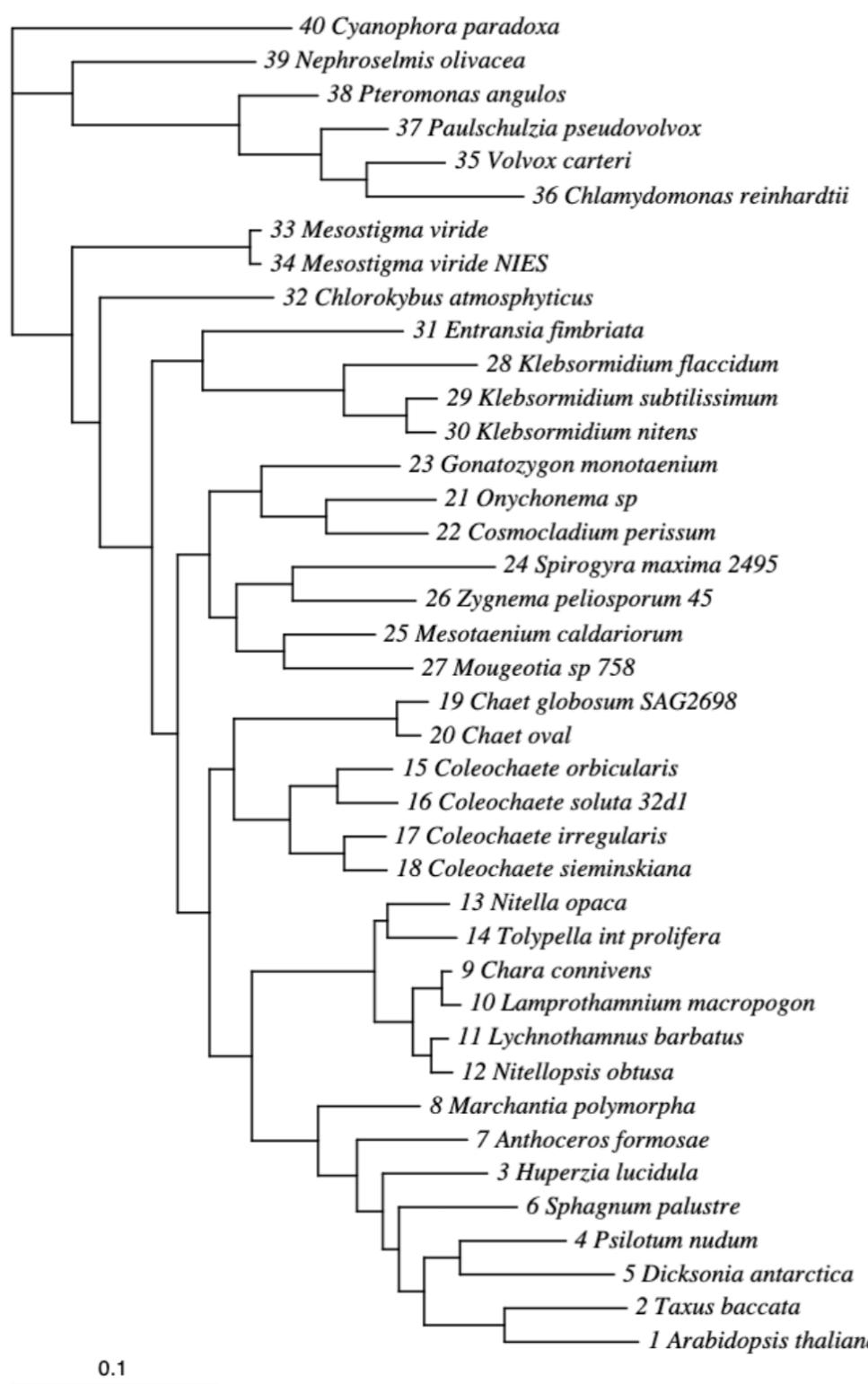


This Gamma prior also has a mode at 3, but has a variance 40 times smaller. Decreasing the variance is tantamount to increasing the strength of the metaphorical rubber band.

Now you (or the likelihood) would have to tug on the parameter fairly hard for it to have a value as large as 4.

This gamma distribution has shape 91.989 and scale 0.032971

# Example: Internal Branch Length Priors

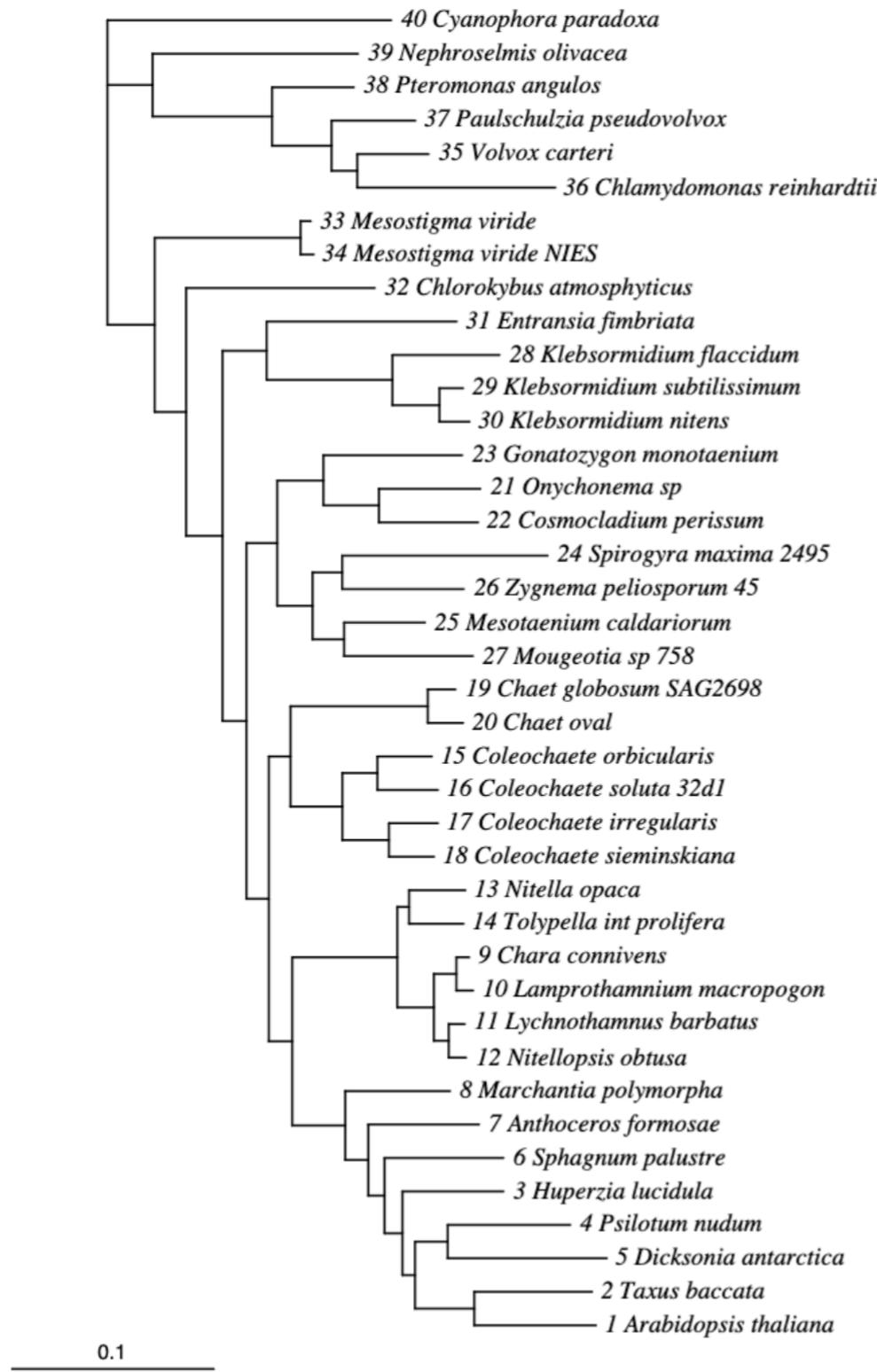


Separate priors applied to internal and external branches

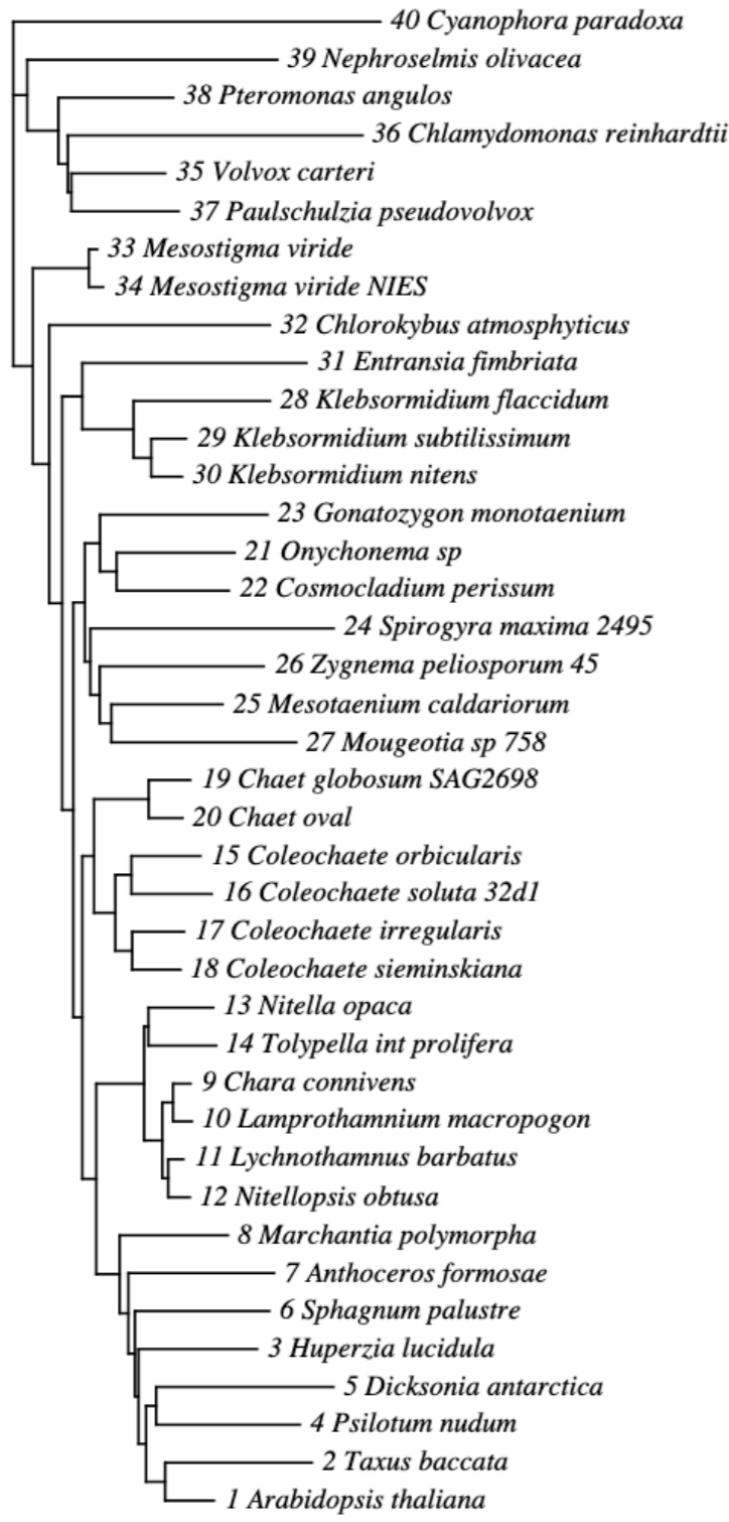
External branch length prior is exponential with mean 0.1

Internal branch length prior is exponential with mean 0.1

This is a reasonably vague internal branch length prior

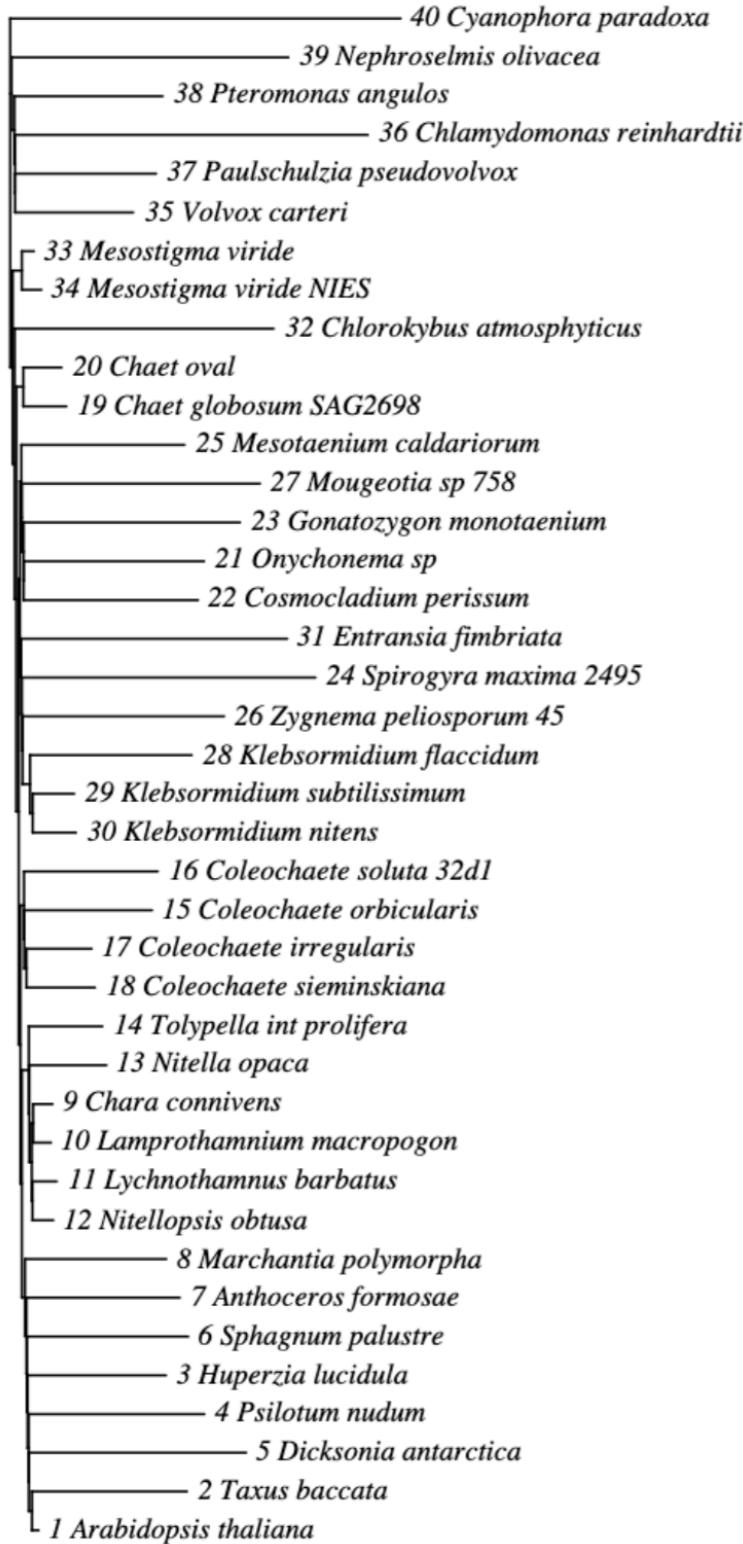


# Internal branch length prior mean 0.001



**Internal branch length prior mean**  
**0.0001**

0.1



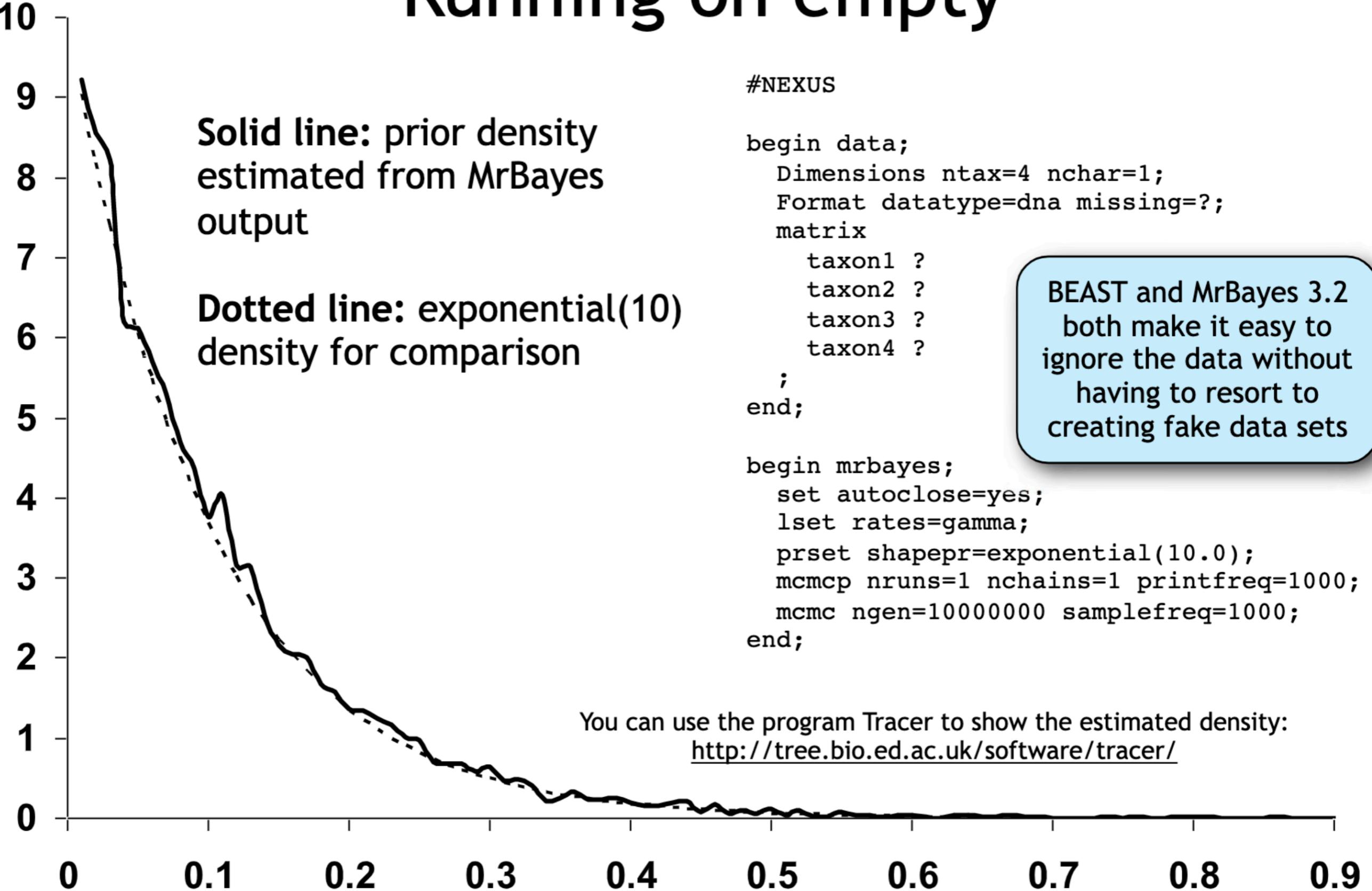
Internal branch length prior mean  
0.00001

0.1

# Running on empty

- Run MCMC without data
- Some programs generate dummy “empty” alignments that can be used
- Why run MCMC on the prior?
  - Check correctness of the software (do the results match the theoretical prior?)
  - Compare prior run with posterior run:
    - If prior and posterior are too similar, then data might contain little information
    - If prior and posterior overlap, but the posterior is more concentrated, then data are informative and prior is reasonable
    - If prior and posterior do not overlap well, then the prior might be misspecified
- CAUTION: do not set a prior to match the posterior! The prior is meant to reflect our knowledge BEFORE the analysis of the data

# Running on empty



# In-class dynamic

- **Time:** 25 minutes
- **Instructions:** Choose a software that does maximum likelihood and follow the tutorial
  - PhyloBase: HAL 1.5 ([github repo](#))
  - MrBayes: [tutorial](#)

and create our own reproducible script. **Bonus points** for paying attention to the MCMC considerations that affect performance in the chosen method

- **Disclaimer:** I have not done the steps ahead of time to make sure that everything runs smoothly so that we can troubleshoot this pipeline together
- **Options for you:**
  1. "I think that I can follow the pipeline by myself or with a small group of peers": you should join the Congregate room
  2. "I think I need more one-on-one help to run the commands": you can stay here in the zoom room